A Constant Percentage Bandwidth Transform For Acoustic Signal Processing

JAMES E. YOUNGBERG

UNIVERSITY OF UTAH

AD A085666

LEVEL II

DTIC ELECTE JUN 1 8 1980 S D C

This document has been approved for public release and sale; its distribution is unlimited.

DDC FILE COPY

JANUARY 1980
UTEC-CSc-80-004
COMPUTER SCIENCE, UNIVERSITY OF UTAH
SALT LAKE CITY, UTAH 84112

80 6 16 18

A CONSTANT PERCENTAGE BANDWIDTH TRANSFORM FOR ACOUSTIC SIGNAL PROCESSING

JAMES E. YOUNGBERG

(14) UTEC-CSC-8Ø-ØØ4

SECURITY CLASSIFICATION OF THIS PAGE (*When Data Entered*)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER UTEC-CS-80-004	2. GOVT ACCESSION NO. AD-A085666	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (*and Subtitle*) (6) A Constant Percentage Bandwidth Transform For Acoustic Signal Processing.		5. TYPE OF REPORT & PERIOD COVERED (9) Technical Report.
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) (10) James E. Youngberg		8. CONTRACT OR GRANT NUMBER(s) (15) N00173-C-79-0045, ARPA Order-33Ø1
9. PERFORMING ORGANIZATION NAME AND ADDRESS University of Utah Computer Science Department Salt Lake City, Utah 84112		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Project: 76-RPA-3301 ARPA Order
11. CONTROLLING OFFICE NAME AND ADDRESS Defense Advanced Research Project Agency (DoD) 1400 Wilson Boulevard Arlington, Virginia 22209		12. REPORT DATE (11) Janu~~ary~~ 1980
		13. NUMBER OF PAGES 110 (12) 122
14. MONITORING AGENCY NAME & ADDRESS(*if different from Controlling Office*) Naval Research Laboratory 4555 Overlook Avenue, S.W. Mail Code 2415-A.M. Washington, D.C.		15. SECURITY CLASS. (*of this report*) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (*of this Report*)

This document has been approved for public release and sale; its distribution is unlimited

17. DISTRIBUTION STATEMENT (*of the abstract entered in Block 20, if different from Report*)

DTIC
ELECTE
JUN 1 8 1980
S D
C

18. SUPPLEMENTARY NOTES

19. KEY WORDS (*Continue on reverse side if necessary and identify by block number*)

Constant-Q, Short-time Fourier transform, rate-change of speech, speech expansion, speech rate compression, phrase unwrapping, perception, filterbank, speech analysis, speech synthesis Fourier transform, auditory analysis, constant percentage bandwidth, constant selectivity, spectral modification.

20. ABSTRACT (*Continue on reverse side if necessary and identify by block number*)

This paper describes a constant percentage bandwidth transform for acoustic signal processing. Such a transform is shown to emulate behavior found in the human auditory system.

A synthesis transformation is developed which, when cascaded with the analysis transformation, provides an analysis-synthesis identity in the absence of spectral modification. over

404949 JOB

The effects of spectral domain modification are described. Principles governing discrete implementation of the transform pair are discussed, and relationships are formalized which specify minimal sample requirements for the spectral domain.

The Constant-Q spectral magnitude and phase functions are discussed, and three main methods are evaluated whereby the spectral phase may be unwrapped.

Finally, the use of the transform pair is discussed in the solution of the perception-related problem of time scale compression and expansion of speech.

Accession For

NTIS GRA&I	☑
DDC TAB	□
Unannounced	□
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or special
A	

A CONSTANT PERCENTAGE BANDWIDTH TRANSFORM
FOR ACOUSTIC SIGNAL PROCESSING

by

James E. Youngberg

January 1980 UTEC-CSc-80-004

This research was supported by the Advanced Research Projects Agency of the Department of Defense under Contract No. Noo173-79-C-0045

ABSTRACT

This paper describes a constant percentage bandwidth transform for acoustic signal processing. Such a transform is shown to emulate behavior found in the human auditory system, making possible both the imitation of peripheral auditory analysis, and processing which is more closely linked to perception than is possible using constant bandwidth analysis.

To enable such processing, a synthesis transformation is developed which, when cascaded with the analysis transformation, provides an analysis-synthesis identity in the absence of spectral modification. Various properties of the transform pair are derived, and a filterbank analogy is used to create a basis for intuitive understanding of the transform's operation and properties.

The effects of spectral domain modification are described and shown to be related to the properties of the analysis window function.

Principles governing discrete implementation of the transform pair are discussed, and relationships are formalized which specify the sampling of the spectral domain. These relationships are shown to depend simultaneously on the analysis window function and the selectivity (or Q) of the analysis. An alternative form of the synthesis is given which facilitates a more nearly optimal logarithmic sampling of the spectral frequency axis. A minimal sampling pattern is given for the spectral domain which has an overall rate equivalent to the rate necessary to sample the constant bandwidth spectral domain.

The nature and computation of the constant-Q spectral magnitude and phase functions is discussed, and three main methods are evaluated whereby the spectral phase may be unwrapped.

Fine resolution constant-Q spectrograms are presented which show clearly the properties of constant-Q analysis applied to speech.

The use of the transform pair is discussed in the solution of the perception-related problem of time scale compression and expansion of speech. Results of this experiment are discussed.

Finally, suggestions for further research and applications are presented.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

The author expresses his appreciation to Dr. Steven F. Boll for guidance and encouragement during the course of this research. Dr. Boll's perceptive suggestions during the presentation of the manuscript are particularly appreciated.

Thanks is also due to James Kajiya, Tracy L. Petersen and especially to H. Ravindra for innumerable discussions and suggestions relating to this work.

The photographic reproductions herein were accomplished by Mike Milochik.

Finally, the author expresses gratitude to his wife, Debby, without whose patient support this project would have been impossible.

CHAPTER 1

INTRODUCTION

1.1 Overview

The usefulness, in signal processing, of transformations which produce spectral representations of temporal or spatial data is rooted in part in the underlying physiological processes which artificial signal processing attempts to imitate or augment. This is particularly true of the transforms used in processing sound signals. Current interest in the short-time Fourier transform, for example, is related to the rough analogy that exists between the short-time spectral domain and the real-time analysis performed by the human inner ear. Because the information obtained from the short-time Fourier transform exists in a format related to the format in which information appears to emerge from the inner ear, intuitive descriptions of signal qualities such as pitch, temporal change, amplitude and harmonic content can easily be related to the properties of the formal mathematical representation. Such a relationship gives insight into both the underlying physiological processes involved, and into artificial processes which may be implemented to affect perception-related changes.

Properties of transformations such as those of the short-time Fourier transform which relate to properties of physical systems determine the appropriateness of such transforms as models. Clearly, as a model's properties more completely conform to the properties of the system which it attempts to emulate, it becomes more useful as a tool for discovery of further system properties, and for duplication and augmentation of processes known to occur within the real system.

An examination of the properties of the short-time Fourier transform as a model of the hearing process will be taken up in the following section, leading to the conclusion, already expressed by several researchers, that it lacks some characteristics essential to the analysis performed in the human auditory system. Section 1.3 then presents preliminary evidence that a constant percentage bandwidth (or constant-Q) transform should more adequately model the human auditory system. Finally, Section 1.4 describes the contribution offered by this work, and outlines the remaining chapters.

1.2 Short-time Fourier Transform Modelling of Human Auditory Signal Analysis

A complete description of the electrical or mechanical analogs proposed by researchers in their attempts to partially account for properties of the peripheral auditory system, is beyond the scope of this work. The subject is treated in numerous references [1,2,3]. In addition, the details of the physiological system or of its analogs are complex, resisting concise mathematical modelling, and hence have not, to date, been useful in solving the usual signal processing problems -- noise removal, parameter extraction and transmission, etc.. A mathematical transformation, on the other hand, can relate more easily to these problems. One such transform, the short-time Fourier transform, the properties of which are well-understood, provides both a forward and a reverse mapping to a domain which resembles the analysis domain of the ear. However, even the most superficial examination of the ear's physiology reveals weaknesses in the short-time Fourier transform as a model of auditory analysis. The transform conforms to the ear-property hypothesized by Helmholtz [4], and later corroborated by Bekesy [5] and others, wherein spatially selective time-limited frequency analysis is performed. It fails, however, to emulate other aspects of the behavior observed by Bekesy. In particular, Bekesy observed that the basilar membrane behaves as a non-uniform or dispersive transmission line such that tones travel a distance inversely proportional to their frequency where they are sensed and then are rapidly attenuated. He further observed that the envelope of a tone traveling the length of the membrane maintains its shape as it moves the 35mm to the apex of the membrane. In other words, the mechanical analysis performed by the inner ear was reported by Bekesy to have a rather low, but constant Q. (The selectivity, Q, of an instrument is a measure of its ability to resolve or respond to a particular frequency component independent of the presence of nearby spectral components. Selectivity is formally defined as the ratio of the center frequency of a response peak to the -3 decibel bandwidth of that peak.) Though it has been extended in accuracy, Bekesy's basic result that frequencies are resolved with roughly constant selectivity at positions logarithmically spaced along the length of the basilar membrane still seems correct. The value for the Q of the ear's analysis, convergently verified by recent experiments, has been specified by Searle [6]. He gives the resolution as roughly one third of an octave (a Q equal to about 4.3). As described in Chapter 2, the short-time Fourier transform behaves as a bank of equally spaced, constant bandwidth filters. Hence, analysis preformed at high frequencies is over-resolved in frequency while that performed at low frequencies may be under resolved. This difficulty in the constant bandwidth short-time Fourier transform has been noted by Callahan [7], who points out that for speech analysis, the window length is a compromise between adequate time resolution at high frequencies and enough frequency resolution at low frequencies. In speech processing schemes where accurate pitch and vocal tract resonance information are required simultaneously, the dilemma is insoluble, and separate pitch extraction is ordinarily necessitated. Clearly, constant-bandwidth analysis fails in this respect as a model of peripheral auditory analysis. The list of other ear phenomena not described (at least not trivially) by the constant bandwidth analysis model includes Tartini's combination tones, Seebuck and Schouten's residue pitch, Sachs and King's two tone suppression and many other phenomena. The above phenomena are complex and have been described by Searle [6] as having second order importance in initial efforts to model the ear via mathematical transformations.

Despite the failure of the short-time Fourier transform to model the essentially constant-Q nature of analysis performed by the ear, its two dimensional nature has proven useful in many applications. Among these are the phase-vocoder [8,9,10], perceptual rate change proposed by Flanagan [1] and recently implemented by Portnoff [11], and various two-dimensional modification experiments involving noise removal, feature isolation and enhancement and bandwidth compression all performed by Callahan [7]. Both Portnoff and Callahan noted the limitations imposed by the constant bandwidths in their experiments, and pointed out the possible advantage inherent to a constant-Q implementation of their systems.

1.3 The Constant-Q Alternative

The notion of constant-Q signal analysis is not new. The analog spectral analyzer has been performing constant percentage bandwidth analysis for decades. That constant-Q analysis could be formalized mathematically was recognized in 1971 by Gambardella [12], who proposed a "multiple filter analyzer integral."

$$F(\omega,t) = \int f(\tau)h(t-\tau,\omega)e^{-j\omega\tau}\,d\tau \qquad (1.1)$$

(Note that integration intervals for all integrals in this work are assumed to be $(-\infty,\infty)$ unless otherwise stated.' This analysis integral is a generalization of the short-time Fourier integral transform in the sense that its analysis window is a function not only of time, but also of analysis frequency. Gambardella pointed out that certain forms of this integral function permit a reverse transform, and that one particular form exhibits constant-Q character.

Related efforts directed at the problem of constant-Q signal analysis have centered attention on the notions of warping the frequency axis and non-uniform sampling of the z-transform. These efforts are reviewed in Chapter 3.

1.4 Contribution and Outline of this Work

The contribution of this work involves formalization of a constant-Q transform and the definition of the properties of the transform as it pertains to acoustic signal processing. Attention has been given to mathematical establishment of both the properties of the forward transform and and reverse transform so that processing which uses the transform may be well-understood.

The effect of spectral domain modification, an important issue where signal processing is the goal of analysis, has been discovered and described. Because, to date, no form of the analysis transform analogous in speed and elegance to the FFT has been found, care has been taken to derive and articulate relationships governing the sampling of the constant-Q spectral domain. A pattern which allows minimal sampling has been described, and an algorithm presented whereby sampled analysis may be achieved at the expense of a complex demodulation and fast convolution on each analysis channel. The nature and computation of the spectral magnitude and phase functions has been discussed, including the problem of spectral phase unwrapping. As an illustration of the use of the transform pair, the perception-related problem of time compression and expansion of speech was solved using the transform,and the performance of the algorithm in this application evaluated.

Chapter 2 contains a discussion of generalized short-time Fourier transform analysis and synthesis, as well as a discussion of the effects of spectral modifications. This material is provided primarily for reference, since many constant-Q concepts are more easily understood by analogy with constant bandwidth concepts.

Chapter 3 then presents the constant-Q transform in a development which parallels that in Chapter 2. One of the family of possible reverse transforms is developed. The effect of constant-Q spectral modification is discussed, and a number of useful transform properties are given.

Chapter 4 handles a collection of topics which do not properly fit into Chapter 3, but which are of practical importance. These include the various implementation issues, such as sampling, filterbank design, computation schemes, and the nature and computation of the constant-Q spectral magnitude and phase functions.

Chapter 5 describes the use of the transform to effect modification of the rate of articulation of speech (not to be confused with scaling the time index). A comparison is made between this and previous work with this problem.

CHAPTER 2

THE SHORT-TIME FOURIER TRANSFORM

2.1 Introduction

The mathematics of the short-time Fourier integral transform and its discrete counterpart have been clearly laid out in a number of standard sources. However, because many of the concepts of the following chapter closely parallel ideas encountered in such a development, a brief review of the continuous forward and reverse transforms, their properties, and the effect of spectral modifications will be presented here. We shall also find it convenient to introduce in this familiar development many of the symbols and terminology used throughout this work.

2.2 Short-time Fourier Analysis

The continuous Fourier integral transform,

$$F(\omega) = \int f(t)e^{-j\omega t}\, dt \tag{2.1}$$

and its inverse,

$$f(t) = \frac{1}{\pi} \int F(\omega)e^{j\omega t}\, d\omega \tag{2.2}$$

have a fundamental limitation: while $F(\omega)$ has infinitesimal frequency resolution, it fails to provide any information about how frequency information varies as a function of time. The concept of spectral information which changes with time does not exist. This limitation is remedied by weighting the time signal in an area of interest using a function which is generally smooth and which has limited non-zero extent. (Such a function is often referred to as an analysis window. The Hann window is a familiar example.) The weighting function imposes, by reason of its non-zero duration, a finite time resolution. A less obvious effect of forcing the Fourier integral transform to operate locally is that the frequency resolution of the transformed signal is no longer infinitesimally fine, since it has been "smeared" by convolution with the Fourier transform of the weighting function. If this resolution-limiting window is allowed to slide along the time axis, as in the short-time Fourier integral transform,

$$F(\omega,t) = \int f(\tau)h(t-\tau)e^{-j\omega\tau}\,d\tau \qquad (2.3)$$

the transform then becomes a function of two variables and yields local (finite resolution) information about the input signal.

2.3 Resolution and Sampling Issues

The terms, "time resolution" and "frequency resolution," used above, require a more formal definition if they are to be useful in discretization of the short-time Fourier transform. Suppose the time and frequency resolutions of constant bandwidth analysis are defined as the time and frequency intervals over which the window function and its Fourier transform are "significant." Clearly this definition involves a degree of ambiguity (and therefore approximation) in the identification of the respective intervals. Hence, a definition is adopted here, equivalent to that used by Allen [13], which allows precise determination of the time and frequency resolutions associated with a window function, h(t). The finite, non-zero extent of h(t) is defined to be the time resolution, T_∞, of the window. If by $H(\omega)$ we denote the Fourier integral transform of h(t), the frequency resolution, F_∞, of the analysis is given by the extent of the principal interval around zero wherein $H(\omega)$ is positive-valued. (The subscripts of T_∞ and F_∞ will become more useful in Chapter 4. They indicate the attenuation of the window or its transform at the edges of the resolution determining interval.)

The time scaling property of the Fourier integral transform [14] guarantees that the product of the time and frequency resolutions will be a constant. Hence, we may write

$$\beta_\infty = T_\infty F_\infty \tag{2.4}$$

where β_∞ is a constant whose value is a consequence of the choice of window function and of the definitions of T_∞ and F_∞. Easily computable values for β_∞ are given in Table A.1 of Appendix A for a few common windows assuming the above-stated definitions of T_∞ and F_∞.

Combined with the celebrated Nyquist sampling theorem, this information is sufficient to permit sampling of the short-time Fourier spectral domain without loss of information. In particular, the density of time samples must be greater than F_∞, and the density of frequency samples must be greater than T_∞. Hence, if the time and frequency sampling intervals are respectively Δt and Δf,

$$\Delta t \leq 2\pi F_\infty^{-1} \qquad (2.5)$$

$$\Delta f \leq (2\pi T_\infty)^{-1} \qquad (2.6)$$

Thus, for instance, a 25.6 millisecond Hann window gives rise to a spectral domain which must be sampled at least every 6.4 milliseconds in time and every 39.0625 Hertz in frequency if information is not to be lost.

(It should be noted that under special conditions [15] restrictions on the analysis window function permit synthesis from undersampled spectral data. In general, however, and where spectral modifications are to be performed, synthesis depends on proper spectral sampling as defined).

The discrete representation of the short-time Fourier transform will not be presented here, but is discussed in several sources [7,10]. The sampling theorem is reviewed primarily because a similar argument will be needed in connection with the constant-Q transform, and because the notion of analysis at discrete frequencies is useful in developing the following analogy.

Suppose the short-time Fourier transform is evaluated at a set of discrete frequencies, $\omega_k = k\Delta f$. If for each k the complex exponential in 2.3 is associated with the input function, the result is recognized to be a convolution (denoted throughout this work by the binary operator, "*").

$$F(\omega_k,t) = f(t)e^{-j\omega_k t} * h(t) \tag{2.7}$$

In this form of the analysis expression, the short-time spectrum at any ω_k is recognized to be a lowpass version of the complex demodulated input signal. A simple change of variables in 2.3 allows the complex exponential to be associated with the window function in the convolution. This results in still another form of 2.3.

$$F(\omega_k,t) = e^{-j\omega_k t} \{f(t) * h(t)e^{j\omega_k t}\} \tag{2.8}$$

Because the various bandpass filters resulting from the complex modulation of h(t) by ω_k form a continuous bank this form of 2.3 has been called the filter bank analogy. It is shown schematically in Figure 2.1.

2.4 Short-time Fourier Synthesis

The nature of the short-time Fourier synthesis integral is suggested by observing that the analysis presents at every instant a frequency-shifted set of contiguous lowpass representations of the original, and

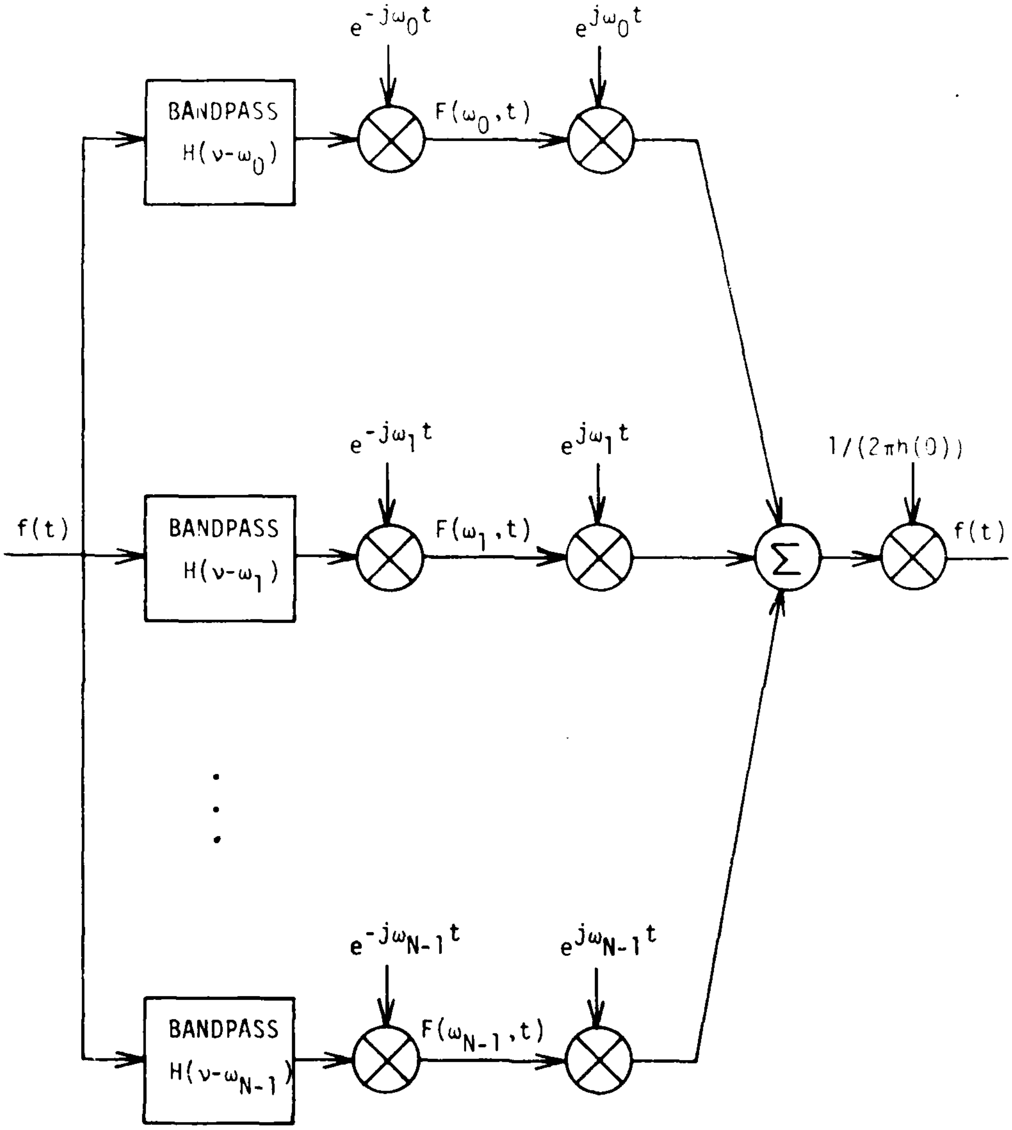


Figure 2.1. Filterbank analogy to short-time Fourier analysis and synthesis. Synthesis via the filterbank summation (FBS) method is shown here. As elsewhere, ν is the Fourier frequency parameter associated with the transformed spectral time axis, and each is an analysis center frequency measured along the ω-axis.

that, re-shifted, these signals should add up to produce a scaled version of the original. This ɫ.oposed synthesis is shown schematically in Figure 2.1. In integral form the synthesis is,

$$f(t) = \frac{1}{2\pi h(0)} \int F(\omega,t)e^{j\omega t}\, d\omega \qquad (2.9)$$

Equation 2.9 is not, however, the most general form for short-time Fourier synthesis, but is in fact a particular case of the more general form given below.

$$f(t) = \frac{1}{2\pi\langle g(t),h(t)\rangle} \int F(\omega,t)g(t-\tau)e^{j\omega t}\, d\omega\, d\tau \qquad (2.10)$$

where <g(t),h(t)> is the inner product,

$$\langle g(t),h(t)\rangle = \langle g,h\rangle = \int g(t)h(t)\, dt \qquad (2.11)$$

of g(t) and h(t), and g(t) is a window function having restrictions similar to those applying to the analysis window, h(t). That this more general form of synthesis provides an analysis-synthesis identity in the absence of spectral modification is shown in Appendix B.

The significance of the existence of a more general form of 2.9 involving a synthesis window will be discussed in Section 2.4. It suffices for now to identify two members of the family of synthesis forms. These two forms result from specifying the synthesis window to be either of the limiting cases, $g(t)=\delta(t)$ or $g(t)=1$. (Here and throughout this work "$\delta(t)$" will represent the Dirac delta "function." Although technically not a function, the Dirac delta provides notational and operational short-cuts when used with care. Its properties and use are described by Papoulis [14] and Lighthill [16]). In the first case, where g(t)=1, 2.10 reduces to

$$f(t) = \iint F(\omega,\tau)e^{j\omega t}\, d\omega\, d\tau \;/\; 2\pi \int h(\tau)\, d\tau \qquad (2.12)$$

If the window area is constrained to equal unity, this expression is recognized as a continuous version of the overlap-add (OLA) synthesis proposed by Allen [13].

The more familiar synthesis expression, 2.9, which for reasons described above is called the filter bank summation (FBS) synthesis is derived from 2.10 by setting $g(t)=\delta(t)$.

2.5 Effect of Spectral Modifications

In some applications, where parameter extraction is the goal, or where complete spectral information is to be transmitted over a noiseless channel, the effect of spectral domain modifications is not important. However, in many applications, modifications to spectral information occur either unintentionally or as a main feature of the processing attempted. In these cases, the effect which spectral domain changes have on the synthesized signal must be understood. As implied in Section 2.3, the OLA and FBS synthesis integrals yield an identity when cascaded with the short-time analysis of 2.3. These synthesis integrals differ, however, in their effect on the mapping between the spectral domain and the time domain if spectral domain signal modifications are allowed. The reason for, and implications of this behavior are the subjects of the present section.

Recall from Section 2.2 that the short-time Fourier transform of a signal has finite time and frequency resolution, given by T_∞ and F_∞. From this fact alone, it is clear that the set of short-time Fourier spectra defined by 2.3 does not, for any given h(t), include every possible complex-valued, two-dimensional function. In other words, the mapping performed by 2.3 from the complex line to the complex plane is not onto. The situation is shown graphically in Figure 2.2 where the shaded area is the subplane reachable from the complex line via the short-time Fourier transform for any particular window, h(t). The family of reverse mappings (called left inverses or retracts) specified by 2.10 maps the whole plane onto the line. If the portion of the plane not reachable from the line could be excluded from our interest, simplicity or computational expediency could dictate our choice of synthesis integral from among the family implied by 2.10. However, because spectral modification often attends spectral domain processing, and because nearly all additive noise as well as many useful modifications map signals outside the subplane, the effect of various retracts on the analysis-synthesis system in the presence of spectral modifications is important.

While many types of spectral modification are possible

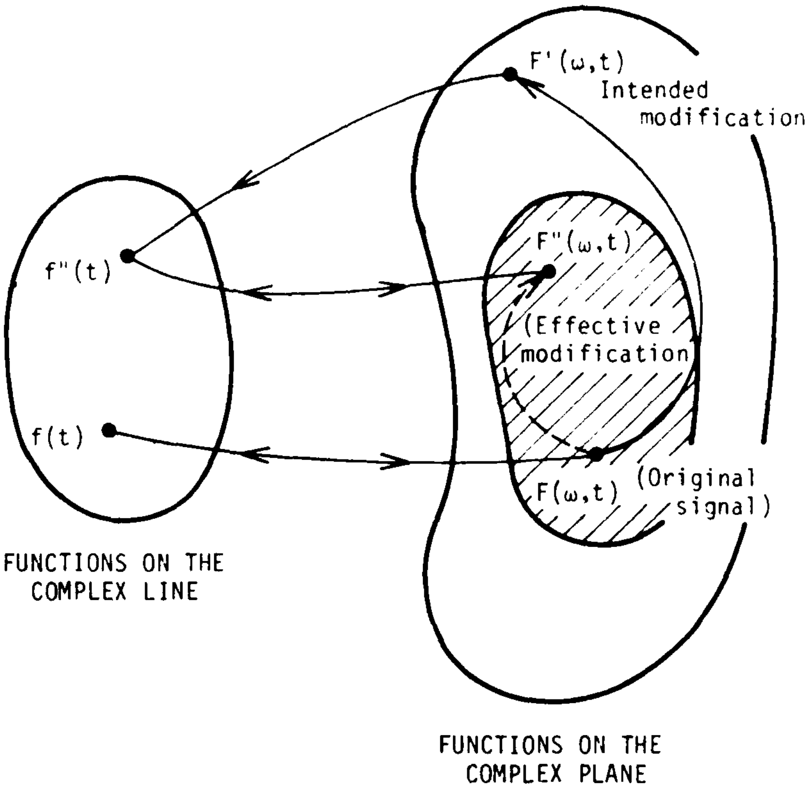


Figure 2.2 The effect of spectral modification. The arrows between the line and the plane indicate mappings available using the short-time (or the constant-Q) forward and reverse transformations. The shaded domain delineates the set of spectra, $F(\omega,t)$, reachable from the time domain. Many spectral modifications map the signals, $F(\omega,t)$, to signals, $F'(\omega,t)$, which lie outside of the shaded region. These "illegal" modifications are mapped to effective "legal" modifications, $F''(\omega,t)$, by synthesis followed by reanalysis.

(convolutional, multiplicative, additive, etc.), the effects to be expected from spectral modification and the method for determining such effects can be conveniently illustrated by considering multiplicative modification. The general relationship between an intended spectral modification and its associated effective modification (see Figure 2.2) may be established for changes of a particular form by substituting 2.3 into a version of 2.10 which reflects the change in question.

Suppose $F(\omega,t)$ in 2.10 is multiplied by the time-varying function, $C(\omega,t)$ as in 2.13.

$$\tilde{f}(t) = \frac{1}{2\pi<g,h>} \iint F(\omega,\tau)C(\omega,\tau)g(t-\tau)e^{j\omega t}\,d\omega\,d\tau \qquad (2.13)$$

Expanding $F(\omega,t)$ in 2.13, reassociating factors, and interchanging the order of integration leads to the following:

$$\tilde{f}(t) = \frac{1}{<g,h>} \iint f(\xi)h(\tau-\xi)g(t-\tau) \cdot \qquad (2.14)$$

$$\frac{1}{2\pi} C(\omega,\tau)e^{j\omega(t-\xi)}\,d\omega\,d\xi\,d\tau$$

In this, the one-dimensional inverse Fourier transform of $C(\omega,t)e^{-j\omega\xi}$, denoted $c(t-\xi,\tau)$, is recognized. Interchanging the order of integration once more yields

$$\tilde{f}(t) = \frac{1}{<g,h>} \int f(\xi) \int C(t-\xi,\tau)g(t-\tau)h(\tau-\xi)\,d\tau\,d\xi \qquad (2.15)$$

If the inner integral is rewritten symbolically as $\hat{c}(t-\xi,\xi)$, 2.15 becomes the superposition integral,

$$\tilde{f}(t) = \frac{1}{\langle g,h\rangle} \int f(\xi)\hat{C}(t-\xi,\xi)\ d\xi \qquad (2.16)$$

If $C(\omega,t)$ is momentarily constrained to be a function of only, 2.15 is seen to be a simple convolution, a result expected from the convolution property of the Fourier transform. If the time dependence is readmitted, interpretation of 2.15 becomes more difficult, but is possible if $g(t)$ is sufficiently constrained.

As examples of the interpretation of 2.15, we will consider the particular cases of $g(t)$ mentioned in Section 2.4. Suppose, for instance, that $g(t)=\delta(t)$ as in FBS synthesis. Then,

$$\hat{C}(t-\xi,\xi) = C(t-\xi,t)h(t-\xi) \qquad (2.17)$$

The intended modification has been time-limited by $h(t)$ (blurred in the frequency domain) but is seen to "take effect" instantaneously in time. This result matches the behavior described by Allen and Rabiner [15] for spectral modifications made prior to FBS synthesis. In the case of OLA synthesis (that is, when $g(t)=1$), 2.16 becomes

$$\hat{C}(t-\xi,\xi) = \int C(t-\xi,\tau)h(\tau-\xi)\ d\tau \qquad (2.18)$$

Note that, in contrast to the FBS result, the intended modification is smeared in time (band-limited by $h(t)$). A more intuitive description of the above effects is taken up in the second section of Appendix B.

The implications of the above results are important. Although the FBS and OLA forms of the general synthesis of 2.9 are most practical, they may not always exhibit desired spectral modification behavior when changes are to be made in the spectral domain. Also, as a result of the fact that time or frequency limiting of intended modifications may occur, smearing in the Fourier transform domain may cause the modification function to extend beyond the limits implied by the spectral time or frequency sampling densities. Hence, care must be taken to sample densely enough in time and frequency to prevent time or frequency aliasing due to spectral modifications.

CHAPTER 3

THE CONSTANT-Q TRANSFORM (CQT)

3.1 Introduction

The purpose of the present chapter is to introduce, in a development similar to that of Chapter 2, the constant-Q transform. In this development, the forward and reverse transforms, their interpretations, the effect of spectral domain signal modifications, and some basic transform properties will be presented. An attempt has been made to describe concepts in terms which facilitate comparison with similar constant bandwidth, short-time Fourier transform concepts.

3.2 Constant-Q Fourier Analysis

Gambardella [17,18] has proposed a generalized short-time Fourier analysis integral for continuous time signals in which the observation window is a function of both the time and frequency parameters of the analysis.

$$F(\omega,t) = \int f(\tau)h(t-\tau,\omega)e^{-j\omega t}\,dt \qquad (3.1)$$

The conventional short-time Fourier integral transform and the standard Fourier integral transform can be considered to be special cases of this transform, obtained from 3.1 when $h(t-\tau,\omega)$ equals $h(t-\tau)$ or when $h(t-\tau,\omega)$ equals unity, respectively. However, as noted by Kajiya [19], a very interesting member of this transform family arises if the window, and therefore the complex transform kernel, is a function of the product of time and frequency. For any analysis frequency, the resulting transform's window length in analysis wavelengths, and therefore the number of cycles of the complex kernel sinusoid, is a global analysis constant. Hence, the measurement of frequency content obtained by integrating the product of the sinusoid and the signal is always the result of estimation over the same length in wavelengths of the frequency in question. This produces the time and frequency resolution effects expected from a constant percentage bandwidth transform. To see more clearly that this is true, the constant-Q transform, given in 3.2,

$$F(\omega,t) = \int f(\tau)h((t-\tau)\omega)e^{-j\omega\tau}\,d\tau \qquad (3.2)$$

will be described in the context of a filterbank analogy similar to that used in Chapter 2 with the short-time Fourier transform.

Suppose the constant-Q spectrum is sampled at a set of frequencies, ω_k , whose spacing does not exceed the upper limit imposed by the analysis resolution of the transform at each frequency. (We shall have more to say about sampling the constant-Q spectrum in Chapter 4). Then, if for each ω_k 3.2 is recognized as a convolution, it may be written as

$$F(\omega_k,t) = f(t)e^{-j\omega t} * h(\omega_k t) \tag{3.3}$$

As in the constant bandwidth case, a simple change of variables in 3.2 leads to an alternative form of the analysis integral.

$$F(\omega_k,t)e^{j\omega_k t} = f(t) * h(\omega_k t)e^{j\omega_k t} \tag{3.4}$$

Fourier transforming both sides of this equation (and invoking the convolution property of the Fourier integral transform, 3.4 can be rewritten as

$$F_\nu(\omega_k,\nu-\omega_k) = F(\nu)H((\nu-\omega_k)/\omega_k)/|\omega_k| \tag{3.5}$$

Here ν is the Fourier frequency parameter, and $F(\nu)$ and $H(\nu)$ are the Fourier integral transforms of f(t) and h(t), respectively. Also, $F_\nu(\omega_k,\nu)$ is the Fourier integral transform of $F(\omega,t)$ with respect to t (again the Fourier frequency parameter is ν). The right hand expression clearly indicates the filterbank behavior of the transform. At each analysis frequency, the input signal is linearly filtered by a basic lowpass filter which has been frequency shifted, then amplitude and frequency scaled by the analysis frequency, ω_k. If each filterbank output is subsequently frequency shifted by $-\omega_k$, the result is that given by 3.4 and shown schematically in Figure 3.1. The difference, then, between the filterbank interpretations of the constant bandwidth and constant-Q transforms lies in the frequency stretching and amplitude scaling of the

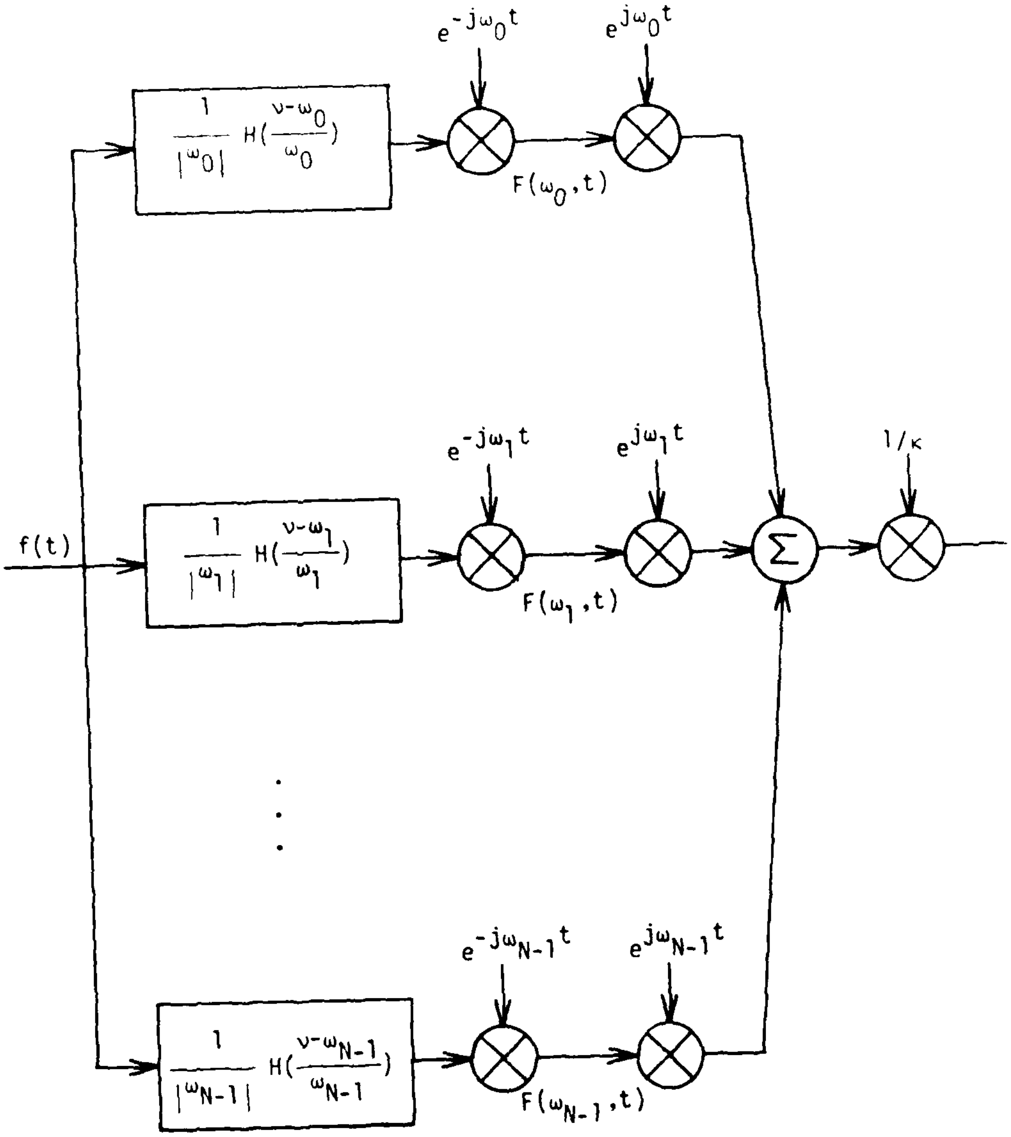
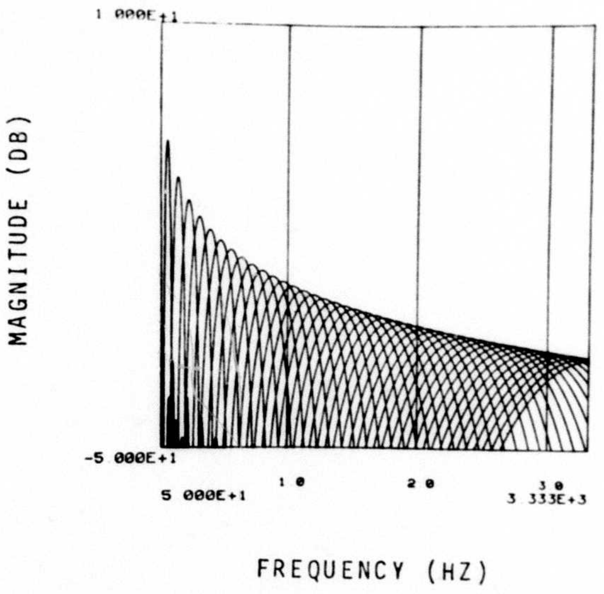


Figure 3.1. Filterbank analogy to constant-Q analysis and synthesis. Filterbank summation (FBS) synthesis, described in the text, is used here.

bandpass filters of the constant-Q filterbank. This difference is crucial however. The frequency resolution, $F_\infty(\omega_k)$, of the kth analysis filter is shown in 3.4 to be directly proportional to its center frequency, ω_k. A bank of such filters is shown in Figure 3.2. On the other hand, the temporal extent, $T_\infty(\omega_k)$, of the kth analysis filter is seen in 3.4 to be inversely proportional to ω_k. Hence, the uncertainty relation which governs time and frequency resolutions for the short-time Fourier integral transform, also governs the resolution of the constant-Q transform, though both resolutions are fixed in the former case. This difference is not unexpected, but is a fundamental stimulus for a study of the constant-Q transform as a model for auditory analysis.

3.3 Other Schemes for Non-uniform Bandwidth Analysis

The importance of the above behavior, although it trivially arises from 3.2, cannot be over-emphasized. Recognizing the fundamental importance of non-uniform frequency analysis, several schemes have appeared in the literature by which Fourier frequency information is sampled at frequency intervals which become wider as frequency increases. A very simple scheme involves sampling the z-transform at non-uniformly spaced points along the unit circle. The recognition that this may cause the highest frequencies to be undersampled suggests the possibility of somehow representing local unsampled



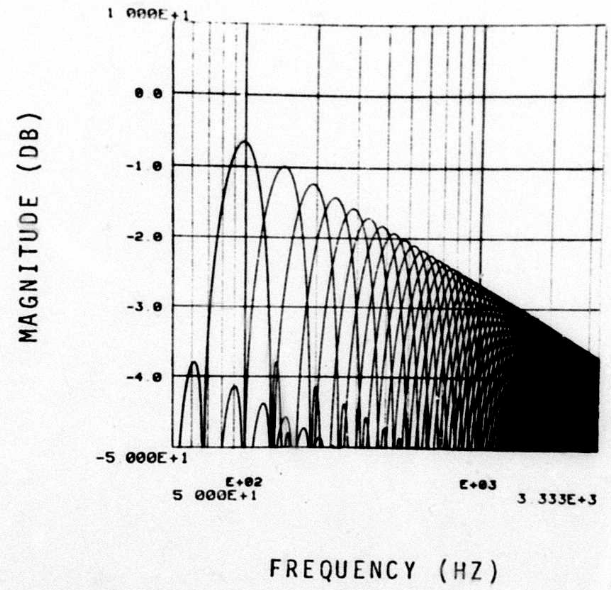


Figure 3.2 Uniformly sampled constant-Q filterbank. This filterbank, undersampled by a factor of five for clarity in presentation, has uniformly spaced, analysis filters. Note that, in spite of the obvious undersampling at low frequencies, the filterbank is heavily oversampled in the high frequencies.

information in the samples which are taken. This is typically attempted by computing a weighted average along the frequency axis of the uniformly sampled short-time Fourier transform in the neighborhood of each new frequency sample. Clearly, the reduced frequency resolution of each sample fails to produce the additional samples required by the implied increase in time resolution, so that unless the short-time Fourier transform is initially oversampled by the amount necessary to produce adequate time resolution after frequency averaging, the information surrendered to the average is lost as surely as if no averaging had been performed. However, with proper attention to sampling issues, this algorithm can be shown capable of producing results equivalent to those formalized in 3.5. Another method is that explained by Oppenheim, Johnson and Steiglitz [20] wherein a sampled input function is passed through a unity magnitude shift-invariant network which produces another sequence whose Fourier transform is related to the Fourier transform of the original sequence by a change of frequency variable. The practical constraint of Fourier transforming the modified sequence using a finite length DFT necessitates windowing the time data. This windowing corresponds to uniform smearing of samples along the new non-uniform frequency axis. Hence, the bandwidth, or frequency resolution of each frequency sample is related to its center frequency and to the frequency domain distortion function. This windowing step gives the method the appearance of constant-Q analysis. However, as will be shown in Chapter 4, a properly sampled constant-Q spectrum has exponentially spaced samples. Unfortunately, the set of frequency axis distortions available to the method does not include a logarithmic transformation. Hence, the method is not truly constant-Q.

Another method, due to Helms [21], approximates the Laplace transform at exponentially spaced frequency intervals, and produces a representation wherein, as frequency increases, the ratio of frequency to bandwidth increases. However, the method is only asymptotically constant-Q.

3.4 Constant-Q Fourier Synthesis

A condition necessary to the general usefulness of any analysis scheme is that the analysis be reversible. Schemes wherein the uncertainty relation is violated are destructive of information and hence analysis performed using these schemes is not reversible. The reversibility of the constant-Q transform will be discussed in this section, and a reverse transform given.

As for the short-time Fourier integral transform (and for the same reason) a true, two-sided inverse does not exist for the constant-Q transform. Rather, a family of left inverses or retracts exist which map the subspace of complex-valued, two-dimensional functions reachable from the complex line via 3.2 back to the complex line.

The nature of one member of this family of reverse mappings is suggested by the observation that in the frequency sampled analog of 3.5 the various filterbank outputs are frequency-shifted outputs of a bank of contiguous bandpass filters. To be sure, the filters are not of uniform width or amplitude; however, the information is all there. This suggests the existence of a continuous synthesis integral of the form

$$f(t) = \frac{1}{k} \int F(\omega,t)e^{j\omega t}\, d\omega \qquad (3.6)$$

wherein the various complex demodulated filterbank outputs of 3.4 are simply remodulated, summed, and normalized by a constant, k. (The value of k will not be defined at this point. It is related to the analysis window.) As an aid to reader intuition, a plausibility argument for this synthesis, also referred to herein as filterbank summation (FBS) synthesis will be given. One way to demonstrate the overall effect of filterbank analysis and synthesis is to compute the frequency response of the entire analysis-synthesis system. Referring to Figure 3.1, the responses of the various filters of the constant-Q filterbank are given by

$$\phi(\omega,\nu) = H((\nu-\omega)/\omega)/|\omega| \qquad (3.7)$$

The overall response of the filterbank may be computed as the sum of the component responses. Such a sum, shown in Figure 3.3 for a discrete-frequency analysis-synthesis system, is expressed symbolically as

$$\Phi(\nu) = \int \phi(\omega,\nu)d\nu = \int H((\nu-\omega)/\omega)/|\omega| \, d\omega \qquad (3.8)$$

A set of conditions sufficient for the existence of this integral is described in Appendix C. For common windows, such as the Hanning window used in this research, the conditions are equivalent to the requirement that the zero-frequency component passed by any filter of a constant-Q filterbank be null. This amounts to a reduction of the set of allowable values for Q to integer multiples of some minimum value.

Given the above existence conditions, the change of variables, $\omega = \alpha\omega$ and $\nu = \alpha\nu$ $(\alpha>0)$, leads to

$$\Phi(\alpha\nu) = \int H((\alpha\nu-\alpha\omega)/\alpha\omega/|\alpha\omega|\alpha \, d\omega \qquad (3.9)$$

which for positive reduces trivially to $\Phi(\nu)$. Since, as shown above, the value of $\Phi(\alpha\nu)$ is independent of α, we must conclude that the value of the filterbank sum, $\Phi(\nu)$, is everywhere a constant. Hence, a properly constructed constant-Q filterbank responds to within a multiplicative constant as an identity system when its outputs are simply summed. The actual value of this multiplicative constant, k (occurring in 3.6), is ordinarily difficult to derive either analytically or numerically. In the author's implementations, empirical determination of the constant was employed.

At this point, it is useful to note that the above

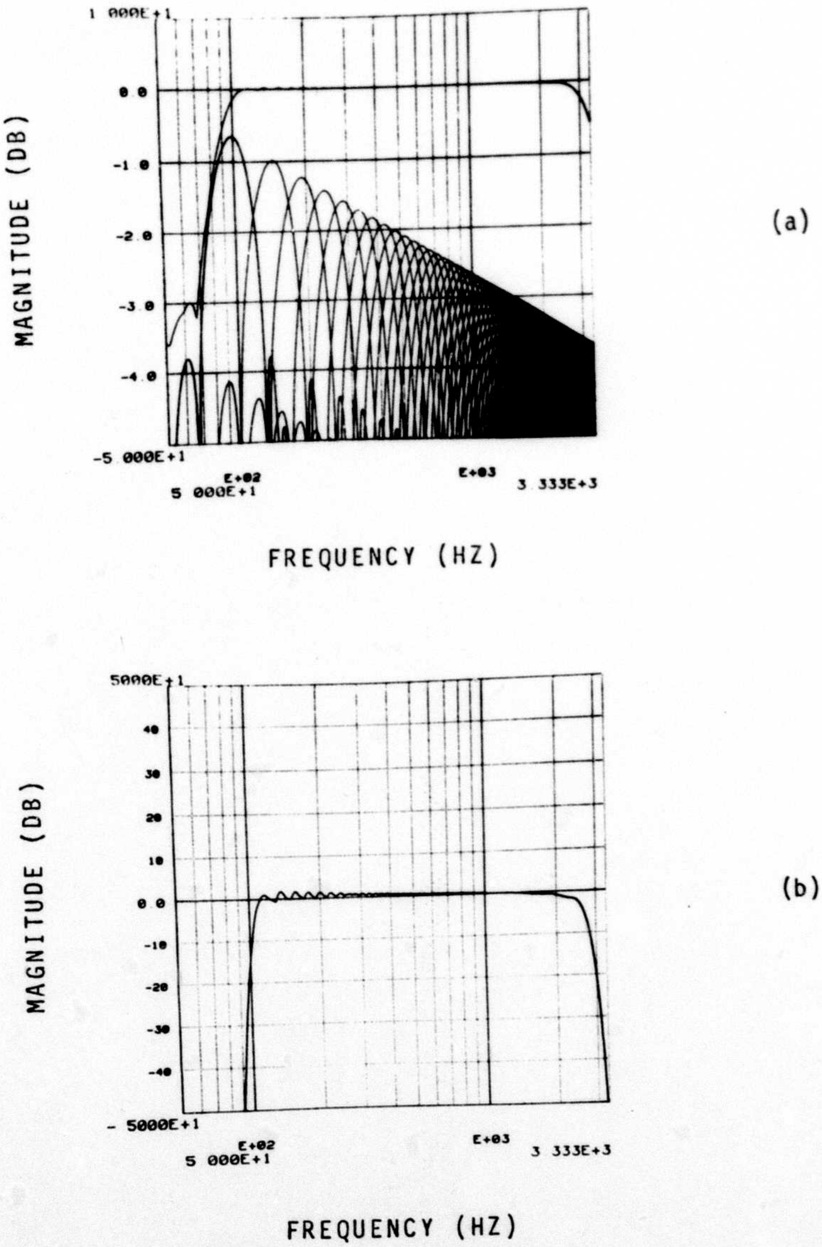


FIGURE 3.3. Discrete-frequency constant-Q filterbank response (uniform sampling). The passband ripple (b), which has a maximum peak-to-peak amplitude of about 0.2 db disappears as frequency increases due to the oversampling shown in (a).

synthesis may also be performed along a logarithmically warped frequency axis if $\omega\phi(\omega,\nu)$ is used as the integrand instead of $\phi(\omega,\nu)$. (Such a filterbank and its sum are shown in Figure 3.4 for a discrete set of frequencies.) The proof is simple, and consists of noting that since

$$d(\log(\omega)) = \frac{1}{\omega} d\omega \tag{3.10}$$

we may rewrite 3.8 as

$$\Phi(\nu) = \int \omega\phi(\nu,\omega)\ d(\log(\omega)) \tag{3.11}$$

This form of synthesis is significant because, as will be seen in Chapter 4, the scheme by which the constant-Q spectral domain is minimally sampled uses exponentially spaced frequency samples.

Another form of the constant-Q synthesis has been proposed by Kajiya [19] in connection with his two-dimensional Mandala transform development. This more general synthesis form is interesting, providing insight into the nature of the above mentioned family of retracts associated with constant-Q analysis. However, the limiting scheme required in the more general synthesis translates less simply into discrete implementation. Hence the analog to short-time FBS synthesis proposed in 3.6 was used in this research.

3.5 Effect of Spectral Modifications

As in the constant bandwidth case, various members of

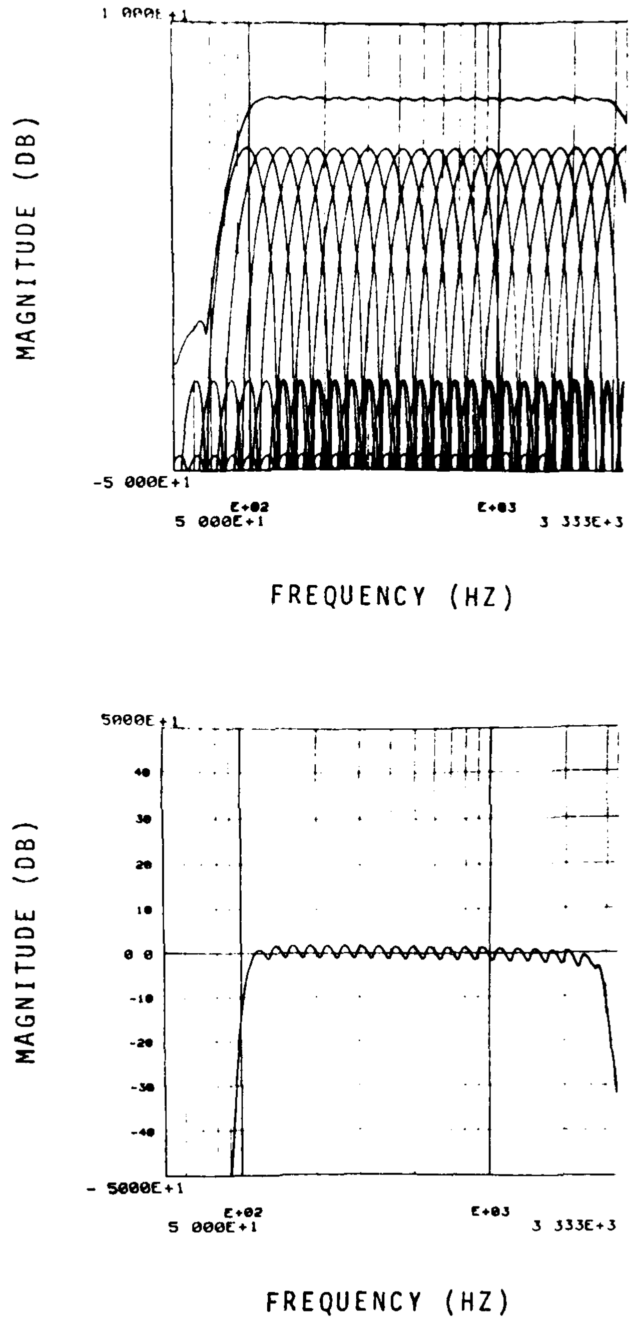


Figure 3.4 Discrete-frequency constant-Q filterbank (exponentially-spaced sampling). Note the linear pre-emphasis which causes the amplitudes of the filters in the filterbank to be uniform. Note also the uniform passband ripple of about 0.4 db peak-to-peak magnitude.

the family of reverse transforms map points in the analysed-signal space which are not reachable via the forward transform back into the signal space differently. The reason for and nature of this fact should be clearly understood before processing is attempted on a spectral domain signal. Unfortunately, the complexity of the forward and reverse transforms makes derivation and interpretation of such information difficult for most modification types. However, an example indicating the technique of such a derivation and showing the interpretation of results will be given here.

Assume that a multiplicative modification which is constant in time is to be applied to a spectral domain signal prior to synthesis via FBS synthesis. Symbolically we write

$$\hat{f}(t) = \frac{1}{k} \int F(\omega,t)G(\omega)e^{j\omega t}\, d\omega \qquad (3.12)$$

$$\hat{f}(t) = \frac{1}{k} \int\int f(t-\tau)h(\omega\tau)e^{-j\omega(t-\tau)}\, d\tau\, G(\omega)e^{j\omega t}\, d\omega \qquad (3.13)$$

$$\hat{f}(t) = \frac{1}{k} \int\int f(t-\tau) \int G(\omega)h(\omega\tau)e^{j\omega\tau}\, d\omega\, d\tau \qquad (3.14)$$

$$\hat{f}(t) = \frac{1}{k} \int f(t-\tau)\hat{g}(\tau)\, d\tau \qquad (3.15)$$

where

$$\hat{g}(\tau) = \int G(\omega)h(\omega\tau)e^{j\omega\tau}\,d\omega \qquad (3.16)$$

Clearly, the effect of such a modification is to filter the input signal using a linear, time-invariant filter. The nature of the effective filter, however, depends on not only the attempted modification function, $G(\omega)$, but as suggested by 3.15 and 3.16, is determined by the analysis window as well. If $\hat{G}(\nu)$ is used to denote the Fourier integral transform representation of $\hat{g}(t)$, the following can be written:

$$\hat{G}(\nu) = \int \hat{g}(\tau)e^{-j\nu\tau}\,d\tau \qquad (3.17)$$

$$\hat{G}(\nu) = \iint G(\omega)h(\omega\tau)e^{j\omega\tau}\,d\omega\; e^{-j\nu\tau}\,d\tau \qquad (3.18)$$

$$\hat{G}(\nu) = \iint G(\omega)H((\nu-\omega)/\omega\ /|\omega|\,d\omega \qquad (3.19)$$

The effective modification is the result of a stylized superposition involving the intended modification and the analysis window function. This operation may be viewed as a weighted sum of the filters in the filterbank. The result of such an operation, even though the weighting function may have arbitrarily fine frequency resolution, is constrained to have frequency resolution which is dictated by the analysis filterbank. Hence, effective modifications are constant-Q versions of intended modifications.

The situation becomes slightly more complicated if the intended modification is allowed to vary as a function of time and frequency. This new multiplicative modifier is denoted below as $C(\omega,t)$.

$$\hat{f}(t) = \frac{1}{k} \iint f(t-\tau)h(\omega\tau)e^{-j\omega(t-\tau)}\, d\tau\, C(\omega,t)e^{j\omega t}\, d\omega \quad (3.20)$$

$$\hat{f}(t) = \frac{1}{k} \int f(t-\tau)\, \hat{C}(\tau,t)\, d\tau \quad (3.21)$$

where

$$\hat{C}(\tau,t) = \int C(\omega,t)h(\omega\tau)e^{j\omega\tau}\, d\omega \quad (3.22)$$

This result parallels the stationary result given above, except that the effective filter is combined by superposition with the input signal. In the frequency domain,

$$\hat{C}(\nu,t) = \int C(\omega,\tau)H((\nu-\omega)/\omega/|\omega|\, d\omega \quad (3.23)$$

Again, the intended modification acts as a weighting function on the various filterbank functions. In 3.23 however, the weighting function is permitted to change instantaneously in time, a result analogous to constant bandwidth FBS synthesis.

3.6 Transform properties

This section states or points out the absence of a few properties of the constant-Q transform which are analogous to the usual Fourier transform properties. In what follows, define

$$f(t) \rightarrow F(\omega,t) \tag{3.24}$$

to be an equivalent statement to equation 3.1.

3.6.1 Linearity Property

If $F_1(\omega,t)$ and $F_2(\omega,t)$ are the constant-Q transforms of $f_1(t)$ and $f_2(t)$, respectively and α_1 , α_2 are two arbitrary constants, then

$$\alpha_1 f_1(t) + \alpha_2 f_2(t) \rightarrow \alpha_1 F_1(\omega,t) + \alpha_2 F_2(\omega,t) \tag{3.25}$$

The proof is a trivial result of the linearity of the integral operator.

3.6.2 Time Scaling Property

If a is a real constant not equal to zero,

$$f(\alpha t) \rightarrow F(\omega/\alpha,\alpha t)/|\alpha| \tag{3.26}$$

To prove this property, assume that $\tilde{F}(\omega,t)$ is the Constant-Q transform (CQT) of $f(\alpha t)$. Then,

$$\tilde{F}(\omega,t) = \int f(\alpha\tau)h((t-\tau)\omega)e^{-j\omega\tau}\, d\tau \tag{3.27}$$

With a change of variables,

$$\tilde{F}(\omega,t) = \int f(\tau)h((\alpha t-\tau)\omega/\alpha)e^{-j\omega t/\alpha}\, d\tau/|\alpha| \tag{3.28}$$

$$\tilde{F}(\omega,t) = F(\omega/\alpha,\alpha t)/|\alpha| \tag{3.29}$$

The absolute value results because, for less than zero, the limits of integration are reversed by the change of variable. Notice that the short-time Fourier integral transform does not share this property.

3.6.3 Time Shifting Property

If t_0 is a real constant,

$$f(t-t_0) \rightarrow e^{-j\omega t_0} F(\omega,t-t_0) \tag{3.30}$$

To prove this property, assume that $\tilde{F}(\omega,t)$ is the CQT of $f(t-t_0)$. Then,

$$\tilde{F}(\omega,t) = \int f(\tau-t_0)h(t-\tau)\omega)e^{-j\omega\tau}\,d\tau \tag{3.31}$$

With a change of variables,

$$\tilde{F}(\omega,t) = e^{-j\omega t_0} \int f(\tau)h(\omega((t-t_0)-\tau))e^{-j\omega\tau}\,d\tau \tag{3.32}$$

$$F(\omega,t) = e^{-j\omega t_0} F(\omega,t-t_0) \tag{3.33}$$

3.6.4 Conjugate Property

If by superscript "*" the operation of complex conjugation is denoted, and if we assume h(t) to be real and even,

$$f^*(t) \rightarrow F^*(-\omega,t) \tag{3.34}$$

The proof is as follows:

$$\tilde{F}(\omega,t) = \int f^*(\tau)h((t-\tau)\omega)e^{-j\omega\tau}\,d\tau \tag{3.35}$$

$$\tilde{F}(\omega,t) = \left(\left(\int f^*(\tau)h((t-\tau)\omega)e^{-j\omega\tau}\,d\tau\right)^*\right)^* \tag{3.36}$$

Changing variables,

$$\tilde{F}(\omega,t) = \left(\int f(\tau)h^{*}((t-\tau)\omega)e^{j\omega\tau}\,d\tau\right)^{*} \qquad (3.37)$$

and with $h^{*}(\omega(t-\tau)) = h(-\omega(t-\tau))$,

$$\tilde{F}(\omega,t) = F^{*}(-\omega,t) \qquad (3.38)$$

3.6.5 Symmetry Properties

If f(t) is real, and if h(t) is real and even, then $F(\omega,t)$ is conjugate symmetric with respect to ω (that is, $F(\omega,t) =-F^{*}(-\omega,t)$).

If f(t) is imaginary, with h(t) as above, then $F(\omega,t)$ is conjugate anti-symmetric with respect to ω (that is, $F(\omega,t) = -F^{*}(-\omega,t)$).

Both of the above follow directly from linearity and the conjugation property.

3.6.6 Other Properties

The independent frequency shifting and scaling properties normally associated with the Fourier integral and short-time Fourier integral transforms do not have simple constant-Q counterparts. Also, the convolution property, absent in the short-time Fourier transform, does not exist for the CQT.

CHAPTER 4

IMPLEMENTATION OF CONSTANT-Q ANALYSIS AND SYNTHESIS

4.1 Introduction

As with the short-time Fourier transform, the usefulness of the constant-Q transform depends on the existence of theory and algorithms which enable it to be applied to discrete data. This need has been met in the first instance by the discrete Fourier transform and by the fast Fourier transform (FFT) algorithm. This chapter considers the problem of computing discrete forward and reverse constant-Q transforms of discrete time data. Issues not involved in a similar discussion of constant bandwidth transform implementation will be shown to arise. The discrete theory and algorithms used in this research will be presented, as will comments concerning other possible implementations.

4.2 Sampling the Constant-Q Spectral Domain

As was pointed put in Section 2.2, the schemes by which the short-time spectral domain may be sampled without loss of information are limited by the analysis window, which imposes a constant time and frequency resolution on the spectral information. The extension of the thinking formalized in Section 2.2 to the problem of sampling the constant-Q spectral domain is complicated by the dependence of $T_\infty(\omega)$ and $F_\infty(\omega)$ on frequency. The solution of this problem necessitates the formalization of some simple ideas. First, define $F_3(\omega)$ to be the -3 decibel frequency extent of the analysis window function, h(t), and β_3 to be the constant product of $T_\infty(\omega)$ and $F_3(\omega)$. Then

$$\beta_3 = T_\infty(\omega)F_3(\omega) \tag{4.1}$$

$$Q = \omega/F_3(\omega) \tag{4.2}$$

Thus, we have an explicit relationship among frequency and the time and frequency resolutions at that frequency.

$$T_\infty(\omega) = \beta_3 Q/\omega \tag{4.3}$$

$$F_\infty(\omega) = \beta_\infty \omega/\beta_3 Q \tag{4.4}$$

Table A.1 of Appendix A lists values of β_3 for a few common window functions. Equations 4.3 and 4.4, combined as in Section 2.2 with the Nyquist theorem, give rise to lower bounds on the local instantaneous sampling densities along the frequency and time axes, respectively. Thus, for example, a Hann window spectral domain whose Q is 3.0 is minimally sampled with a frequency interval of 173 Hertz and a time interval of 1.4382 milliseconds at 1000 Hertz. The same domain at 50 hertz must be sampled at least every 8.6914 Hertz in frequency, but only 28.7641 milliseconds in time. In general, if $\Delta t(\omega)$ and $\Delta f(\omega)$ are respectively the time and frequency sampling intervals at ω,

$$\Delta t(\omega) \leq 2\pi/F_{\infty}(\omega) = 2\pi\beta_3 Q/\beta_{\infty}\omega \qquad (4.5)$$

$$\Delta f(\omega) \leq 1/2\pi T_{\infty}(\omega) = \omega/2\pi\beta_3 Q \qquad (4.6)$$

4.3 Design of a Constant-Q Filterbank

Because to date no useful analog to the DFT has been discovered for the constant-Q transform, the sampling and resolution information described in Section 4.2 must be used to design an analysis algorithm which ultimately invokes discrete convolution to simulate at selected frequencies the action of a constant-Q filterbank. The character of an algorithm which correctly performs this analysis is the subject of the present section.

As implied by 4.5 and 4.6, minimal sampling along both the time and frequency axes is performed according to a non-uniform scheme. The problem of determining at what points the continuous spectral domain ought to be sampled is simplified by considering the domain to be the output of an *analog* filterbank. It is then necessary to specify individual band center frequencies and bandwidths. Figure 3.4 shows a portion of an idealized filterbank which minimally samples the frequency dimension of the constant-Q spectral domain. The relationship between adjacent band center frequencies is established by the following observation, given that the window function, h(t), is real (the demodulated filter magnitude must be even):

$$\omega_{k+1}-\omega_k = \frac{\Delta f(\omega_k)}{2} + \frac{\Delta f(\omega_{k+1})}{2} \quad (4.7)$$

(See Figure 4.1.) By substituting the expression for the instantaneous frequency sampling interval given in 4.6 and by simple algebraic manipulation, the ratio, R, of adjacent band center frequencies may be determined.

$$R = \frac{\omega_{k+1}}{\omega_k} = \frac{2Q\beta_3+1}{2Q\beta_3-1} \quad (4.8)$$

This ratio, along with the location of any band in a constant-Q filterbank, determines the location of any other band as follows.

$$\omega_k = \omega_{k-n}R^n \quad (4.9)$$

Hence, an analog filterbank which performs constant-Q analysis using a minimal set of bands is completely specified by defining the basic window or filter function (from which β_3 is determined), the analysis Q, the total analysis bandwidth, and the center frequency of any analysis band.

4.4 Implementation Details

Until this point, the discussion of *sampling* has assumed continuous time signals and a finite set of analog filters. Thus, only the frequency dimension has been discretized. Discrete-tim_ implementation of the above-specified filterbank is straightforward, requiring application of digital filter design and biplexed [22], fast convolution [23]. Careful attention must be given to

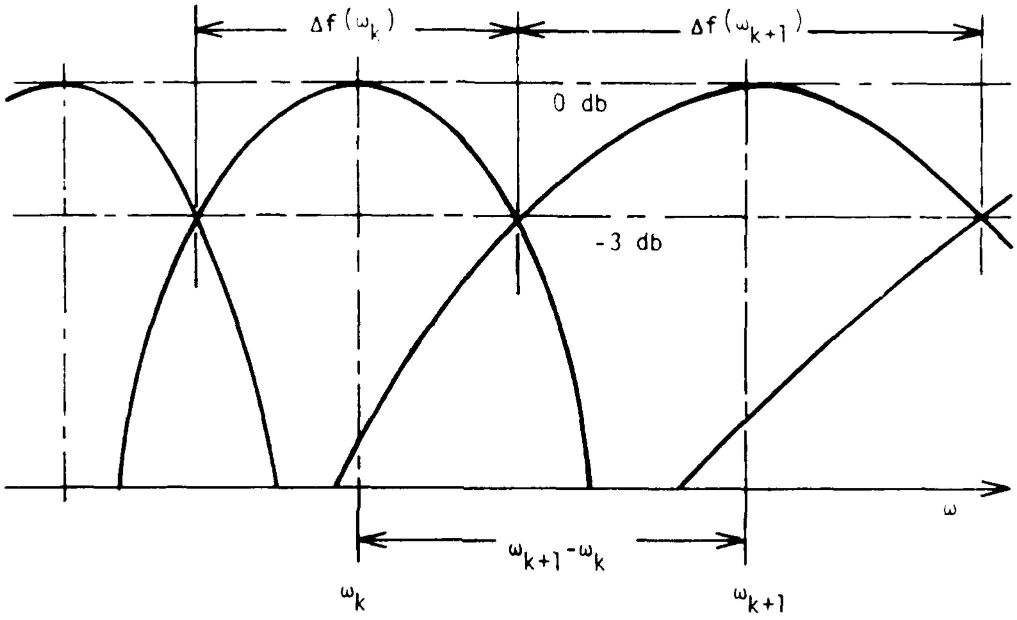


Figure 4.1. Relationships among adjacent filters in a minimally-sampled constant-Q filterbank. This development assumes real analysis window functions (which have even frequency-domain magnitudes).

issues such as elimination of differences in the linear phase introduced by the complex modulation of the various analysis filters. The maximum interval over which the output of the kth band of the filterbank may be sampled is that given by

$$\Delta t(\omega_k) = 2\pi\beta_3 Q/\beta_\infty\omega_k \qquad (4.10)$$

In practice, all analysis channels may be designed to operate at a common sampling frequency which is greater than or equal to the total analysis bandwidth. Hence, for analysis of a segment which has been bandlimited to $2\omega_h$ and sampled at ω_h/π, the total computational expense is equal to the sum of the costs of the individual complex demodulations and fast convolutions for each analysis band.

Of course, this uncomplicated implementation, shown in Figure 4.2a, is wasteful of computational resource. That this is so is made obvious by comparing the total signal bandwidth, $2\omega_h$, at which analysis is performed, with the bandwidths of the highest and the lowest analysis output bands, ω_N and ω_ϕ, given typical analysis parameters. With a third octave Hann filterbank whose highest channel is centered at ω_N and whose lowest channel is centered at $.01\omega_N$, the portion of computation performed unnecessarily varies between 54% at the highest channel and 99% in the lowest, with the average waste equal to 88%. Much of this unnecessary computation could be eliminated by bandlimited resampling of the individual complex channel signals after complex demodulation and prior to (or possibly as a part of) lowpass filtering. Rabiner and Crochiere [24] have described an efficient algorithm for performing bandlimited sampling rate reduction. Their method is a multistage extension of the technique described by Schafer and Rabiner [25] wherein a rate change by a rational factor is accomplished by using a single operation which views a rate change as a cascaded interpolation and decimation, and takes advantage of the bandlimiting by computing only the necessary output points, and by avoiding multiplications involving zeros in the input. The multistage technique cascades such optimal interpolators and decimators to achieve large efficiency improvements, while preserving linear phase and reducing much of the finite arithmetic error associated with the less efficient canonical schemes. Moreover, most of the advantage of the multistage algorithm is gained with only two stages of interpolation and two stages of decimation, all of which may be automatically designed and implemented. Another significant characteristic of analysis using such a system is that the filtering step in analysis may be performed using a single filter rather than the several individual analysis filters suggested by equation 3.5. This scheme for constant-Q analysis is presented in Figure 4.2b. As in the straightforward implementation, attention must be given to the correction of phase disparities which may develop among the analysis channels as the result of non-zero phase

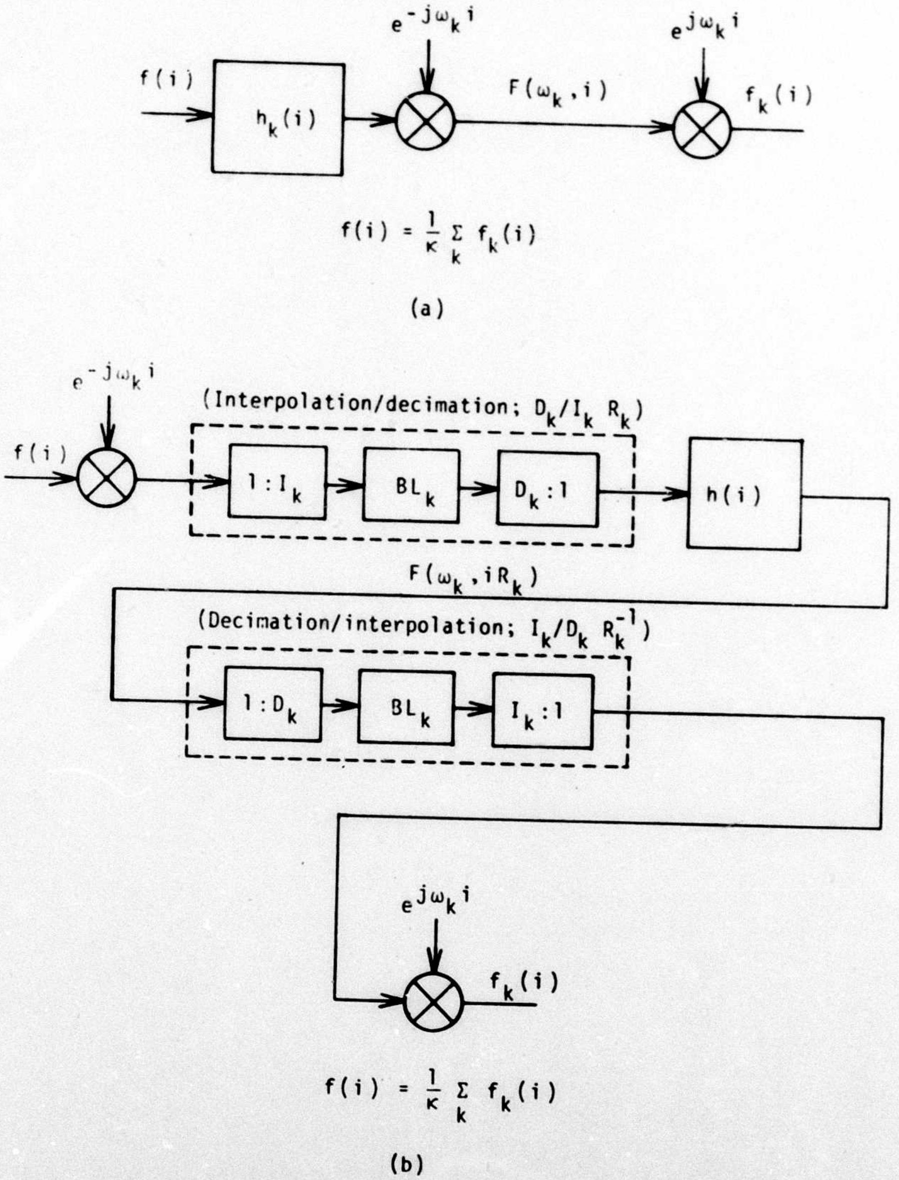


Figure 4.2. Alternative implementations of constant-Q analysis and synthesis. A uniform time-sampling approach is shown in (a), while the implementation suggested in (b) allows minimal time sampling.

resampling filters or because of a non-zero phase analysis filter which is operating on signals of differing sampling rates.

An alternative computational algorithm recently suggested by Kates [26], in connection with perception-related analysis of loudspeaker performance, achieves computational elegance at the cost of the necessity of uniform sampling in both time and frequency. Kates's simplification consists of the restriction that window functions, h(t), be decaying exponentials. If we define h(t) as,

$$h(t) = \begin{cases} e^{-\mu t} & t \geq 0 \\ 0 & t < 0 \end{cases} \qquad (4.11)$$

where μ is a positive, real constant, then 3.2 may be written as

$$F(\omega,t) = \int f(\tau) e^{-\mu\omega(t-\tau)} e^{-j\omega\tau}\, d\tau \qquad (4.12)$$

For any analysis time, t_0, it can be shown that 4.12 is equivalent to

$$F(\omega,t_0) = e^{-j\omega t_0} \int_{-\infty}^{0} f_{t_0}(t)\, e^{-t\omega(j+\mu)}\, dt \qquad (4.13)$$

where $f_{t_0}(t) = f(t+t_0)$. This is easily discretized by periodically sampling $f_{t_0}(t)$ from t=0 at a rate at least equal to its bandwidth in Hertz. Then,

$$P_{cbw} = e^{j\omega t_0} F(e^{j\omega}, t_0) = \sum_{n=0}^{\infty} f_{t_0}(n) e^{-n\omega(j+\mu)} \qquad (4.14)$$

Notice that the z-plane has been evaluated, not along the unit circle, but along the spiral given by $z=e^{\mu\omega}e^{j\omega}$. This particular form of the z-transform is implementable using the chirp z-transform algorithm [27].

Still other implementation schemes are possible, such as that currently being investigated by Tracy L. Petersen [28] employing IIR analysis filters, and the use of charge-coupled dev'ce (CCD) technology which promises significant computation speed improvements for a limited set of applications.

4.5 Minimum Overall Sampling Rate

The non-uniform minimal sampling scheme described in Section 4.4, along with the notion presented in Section 1.2, that constant-Q analysis resembles more clearly than short-time Fourier analysis the dissection of sound performed by the human ear, suggests the possibility of a difference in overall sampling rates needed to represent the respective domains. The overall rate for either spectral domain is easily derived as the sum of the individual rates over all the channels. For constant bandwidth analysis this overall rate, P_{cbw}, is

$$P_{cbw} = \sum_{k=1}^{N} 2p_{cbw} = 2Np_{cbw} \qquad (4.15)$$

where p_{cbw} is the rate for each channel. Thus,

$$P_{cbw} = \frac{2\omega_N}{2\pi\Delta f} \cdot \frac{1}{\Delta t} = \frac{2\omega_N}{2\pi} T_\infty F_\infty = \frac{\beta_\infty}{\pi} \omega_N \qquad (4.16)$$

For constant-Q analysis we sum over a number of channels which grows without bound as zero-frequency is approached. The overall rate, P_{cq}, for constant-Q analysis is

$$P_{cq} = 2p_{cq}(\omega_k) \qquad (4.17)$$

From equation 4.5

$$P_{cq} = 2 \sum_{k=0}^{\infty} \frac{1}{\Delta t(\omega_k)} = 2 \sum_{k=0}^{\infty} \frac{\beta_\infty \omega_k}{2\pi\beta_3 Q} \qquad (4.18)$$

Then from equation 4.9

$$P_{cq} = \frac{\beta_\infty}{\pi\beta_3 Q} \sum_{k=0}^{\infty} \omega_N R^{-k} \qquad (4.19)$$

Noting that R must be greater than one, the geometric sum may be evaluated. If the 4.8 expression for R is then substituted,

$$P_{cq} = \frac{\beta_\infty \omega_N}{\pi} \frac{2\beta_3 Q+1}{2\beta_3 Q} \qquad (4.20)$$

For the values of Q ($3 < Q < 20$) and β_3 ($\beta = 1.4$) used in this research, the latter factor is about 1.1. Hence,

$$P_{cq} \approx P_{cbw} \qquad (4.21)$$

In practice, where the portion of the band near to zero frequency is not sampled, this disparity decreases.

4.6 Computation of Constant-Q Spectral Magnitude and Phase

The final topic to be considered in this chapter on the theory affecting implementation is the computation of the constant-Q spectral magnitude and phase functions from the analysis output. That the spectral domain is complex valued is obvious from equation 3.3 which shows $F(\omega,t)$ for any analysis frequency, $\omega=\omega_k$, to be the output of a linear system whose input is a complex demodulated real signal. If we denote the real and the imaginary parts of the constant-Q spectral domain as follows,

$$F(\omega_k,t) = F_R(\omega_k,t)+jF_I(\omega_k,t) \tag{4.22}$$

then the spectral magnitude and principal value phase functions are computed in the usual manner as

$$M(\omega_k,t) = \{F(\omega_k,t)F^*(\omega_k,t)\}^{\frac{1}{2}} \tag{4.23}$$

$$\theta(\omega_k,t) = \begin{cases} \tan^{-1}(F_I(\omega_k,t)/F_R(\omega_k,t)) & F_R\geq 0,\ \text{all } F_I \\ \tan^{-1}(F_I(\omega_k,t)/F_R(\omega_k,t))+\pi & F_R<0,\ F_I>0 \\ \tan^{-1}(F_I(\omega_k,t)/F_R(\omega_k,t))-\pi & F_R<0,\ F_I<0 \end{cases} \tag{4.24}$$

It should be noted that the above non-linear operations inevitably produce signals which are not band-limited when bandwidth is measured in terms of conventional rectangular width. Hence, it is possible that magnitude and phase functions, computed from adequately-sampled complex data, $F(\omega_k,t)$, could be undersampled. It is also true, however, that the undersampling of the magnitude is not serious. Undersampled areas of the waveform inevitably occur in the magnitude troughs created where either the real or the imaginary part changed sign. Hence, such low energy areas contribute but a small fraction of the total spectral mass. The phase function, on the other hand, when either the real or the imaginary part becomes very small, experiences rapid movements and even discontinuities of π, regardless of efforts to bandlimit it by oversampling. The problem of correctly estimating the spectral phase function is further complicated by "leakage" through the side lobes of the analysis filter, and by finite arithmetic error. These problems, compounded by the fact that only the principal value or wrapped phase is directly computable from the real and imaginary parts, make the estimation of the spectral phase function difficult, particularly in the broadband high frequency analysis channels.

A number of techniques for the estimation of the sampled spectral phase function were investigated during the course of this research. Three general techniques, labelled Methods I, II and III, are outlined below.

Method I for constant-Q spectral phase unwrapping circumvents the necessity of removing the 2π jumps inherent to the principal value inverse tangent function by directly estimating and integrating the time derivative of the phase. The phase derivative is estimated [1] using the property that

$$\dot{\theta}(\omega_k,t) = \frac{d}{dt}\tan^{-1}(F_I/F_R) \qquad (4.25)$$

$$\dot{\theta}(\omega_k,t) = \frac{F_R\dot{F}_I-F_I\dot{F}_R}{F_R^{\,2}+F_I^{\,2}} \qquad (4.26)$$

where $\dot{\theta}(\omega_k,t)$ represents the time derivative of the true unwrapped phase. $F_R=F_R(\omega_k,t)$ and $F_I=F_I(\omega_k,t)$, and $\dot{F}_R$ and $\dot{F}_I$ are the respective time derivatives. Then,

$$\theta(\omega_k,t) = \int_0^t \theta(\omega_k,t)\,dt \qquad (4.27)$$

The difficulty with this otherwise elegant scheme is that in a sampled implementation the derivative and integral can be computationally expensive procedures which are, at best, subject to the limitations of finite arithmetic. Hence, the discrete integral may drift from its true value. To minimize this effect, second degree interpolators were used to estimate the sampled derivative and integral functions as follows:

$$\dot{F}_R(i) = (F_R(i-2)-8F_R(i-1)+8F_R(i+1)-F_R(i+2))/12 \qquad (4.28)$$

$$\dot{F}_I(i) = (F_I(i-2)-8F_I(i-1)+8F_I(i+1)-F_I(i+2))/12 \qquad (4.29)$$

$$\dot{\theta}_I(i) = \theta_I(i-1)+(5\dot{\theta}_I(i-1)+8\dot{\theta}_I(i)-\dot{\theta}_I(i+1))/12 \qquad (4.30)$$

where $F_R(i)=F_R(\omega_k,i\Delta t(\omega_k))$ and $F_I(i)=F_I(\omega_k,i\Delta t(\omega_k))$, and where,

$$\dot{\theta}_I(i) = \frac{F_R(i)\dot{F}_I(i)-F_I(i)\dot{F}_R(i)}{F_R^{\,2}(i)+F_I^{\,2}(i)} \qquad (4.31)$$

A second general method, Method II, due to Portnoff

[11], assumes the accuracy of the discrete phase difference as an estimate of instantaneous frequency. Given this assumption we may argue that because each analysis channel signal is bandlimited, the instantaneous frequency of each channel must be similarly bandlimited. Thus, if a channel is sampled at at least twice the rate implied by its bandwidth, the values of the true phase difference must fall in the interval, $(-\pi/2,\pi/2)$. To unwrap phase under these assumptions, we simply add or subtract integer multiples of π to the wrapped phase difference until this condition is met, and then calculate the unwrapped phase as the running sum of the corrected phase differences. An equivalent form of this method, which avoids the accumulation of arithmetic error in the sum is given by

$$\theta_{II}(i) = \theta_p(i)+\pi\left\lfloor\frac{\theta_{II}(i-1)-\theta_p(i)}{\pi}+\frac{1}{2}\right\rfloor \qquad (4.32)$$

where $\theta_p(i)$ and $\theta_{II}(i)$ are respectively the wrapped phase and the estimate of the true, unwrapped phase for a given channel, and $\lfloor x \rfloor$ indicates the floor of x (the largest whole integer in x). The difficulty with this scheme lies in the initial assumption which, for the broadband high frequency constant-Q channels, fails frequently.

Still another method which was investigated, Method III, utilizes the estimate of the phase derivative as computed in Method I, in conjunction with the knowledge that the true phase can differ from the unwrapped phase only by integer multiples of π. Hence, to estimate the phase at any point, the phase derivative is estimated at that point and integrated to produce a phase estimate from which a phase difference may be computed. The estimated phase difference is then added to the phase estimate of the previous point (assumed now to be correct). This new value is then forced to the nearest value which differs from the wrapped phase by an integer multiple of π. Formally, this method is expressed as,

$$\theta_{III}(i) = \theta_p(i)+\pi\left\lfloor \frac{\theta_{III}(i-1)+\theta_I(i)-\theta_I(i-1)-\theta_p(i)}{\pi}+\frac{1}{2}\right\rfloor \quad (4.33)$$

where $\theta_p(i)$ is the wrapped phase, $\theta_I(i)$ the Method I integrated phase derivative estimate, and $\theta_{III}(i)$ the Method III phase estimate. Clearly, the sources of error in this method are the inaccuracy in estimating the phase derivative function, and the integration of the phase derivative across the interval between the previous point and the current point. The cumulative integration errors inherent to Method I do not, however, occur in this method.

Variations of the three methods outlined above were all found to perform imperfectly. The assumption made in Method II, while excellent for narrow band low frequency channels, was inadequate for high channels. The reason for this is shown in Figure 4.3, which is a histogram of phase differences computed on the output of Method I for a high and a low channel (both channels were oversampled by a factor of three). A signal synthesized after unwrapping using Method I contained less error-induced noise than the

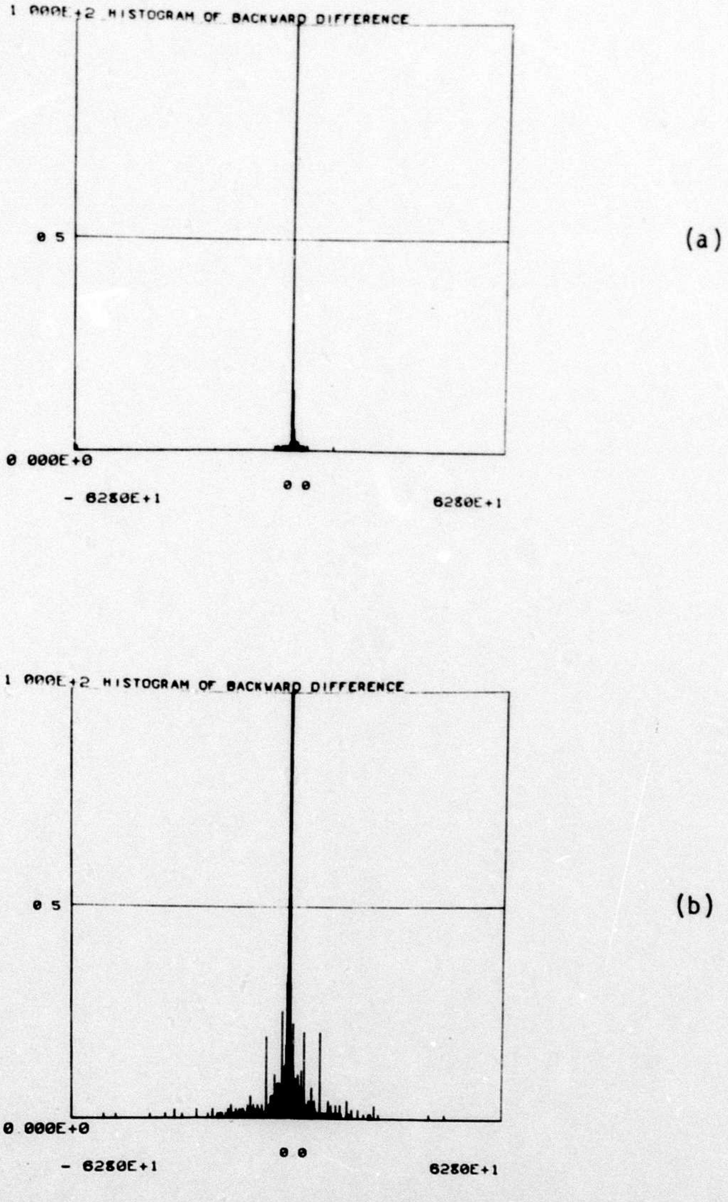


Figure 4.3. Channel phase-difference histograms. To produce these histograms, a signal was upsampled by a factor of three, analyzed with Q=11.5, and unwrapped using Method I. Discrete phase differences were then computed. In (a), computed from a low-frequency channel, phase differences are concentrated in the region immediately surrounding zero frequency. In (b), however, phase differences extend beyond the interval argued as justification for Method II. The histogram (b) was computed from a high-frequency channel.

signal synthesized after unwrapping via either Method II or Method III. Method III, as might be expected, produced fewer unwrapping errors than Method II, resulting in the introduction of slightly less error-induced noise than was observed in the use of the Method II unwrapper. Apparently, the ear is more tolerant of gradual phase drift than it is of sudden jumps. In the author's experience, such gradual changes were noticeable only as a low-level random modulation of background hiss. Sudden phase displacements, on the other hand, gave rise to a "gurbling" noise, the subjective intensity of which increased with the greater number of errors committed by Method II. This noise was occasionally large enough to partially mask low energy stops or fricatives.

4.7 Some Additional Notes on Phase

It should be noted that, for most applications, it is unnecessary to estimate the constant-Q phase function. The spectral phase is important in this research for two reasons. First, its correct estimation is critical to the solution of the rate modification problem formulated in Chapter 5. The second reason for interest in the constant-Q spectral phase involves its role in conveying perceptually important signal information. Callahan [7] and others have noted that in speech processing, depending on the analysis window resolution, the short-time spectral phase contains mostly excitation information, while the spectral magnitude is dominated by vocal tract or formant information. Analogous behavior was found to occur in the case of the constant-Q analysis. This condition was tested by observing the results of syntheses performed using magnitude only (spectral phase set to zero) and phase only (magnitude set to unity). In listening tests, it was observed that, for analysis selectivities substantially larger than the ears' (Q=11.5 for example), vocal tract information was concentrated principally in the magnitude with a very small portion appearing in the signal reconstructed from phase only. Both large Q syntheses contained easily distinguishable pitch information, but the phase signal was found to be the dominant carrier of excitation information. Lowering the analysis Q to be comparable to or less than the ears' selectivity increased the dominance of vocal tract information in both the magnitude and phase only reconstructions. Hence, for low Q, the magnitude-only synthesis contained no pitch information, while the phase synthesis contained both vocal tract and excitation information. These results agree with the observations made by both Flanagan [1] and Callahan [7] that narrow bandwidth channels in a filterbank force excitation information into the phase signal, while wider bandwidth channels can convey this information via the magnitude. That the magnitude contains much excitation information for low Q is shown in Figure 4.4. (In this and other spectrograms appearing in this work, light intensity

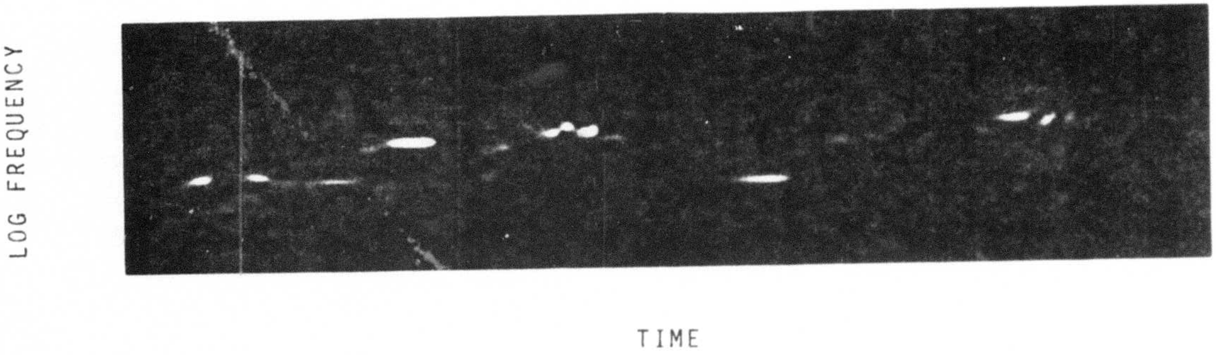


Figure 4.4 Constant-Q spectrogram showing the simultaneous presence of pitch and vocal tract information in the spectral magnitude. This spectrogram was computed with an analysis Q equal to 8.0 on the sentence, "We were away a year ago," spoken by a male and digitized at 10 KHz. The frequencey axis was sampled using an exponentially spaced, pre-emphasized filterbank of 34 channels similar to the filterbank shown in Figure 3.4 In this display, light intensity is proportional to spectral magnitude. Each spectrogram consists of two sections which should be contiguous in time, and which are scaled along the time axis at intervals of 0.1 seconds.

is proportional to the preemphasized spectral magnitude, $|\omega F(\omega,t)|$. Thus, the bright areas indicate the presence of spectral power, while the darker background areas occur where less activity is present.) Note the strong horizontal line (and its harmonics) corresponding to pitch, as well as the perioaic pitch-related variation in the energy of the higher frequency structures often referred to as formants.

CHAPTER 5

TEMPORAL AND HARMONIC SCALING

5.1 Introduction and Background

The auditory system is, next to the visual system, the broadest bandwidth channel available for communication with the human mind. Indeed, when comprehension is used as a measure, evidence [29] suggests that the auditory channel may exceed the visual channel in its usefulness in information transfer. Yet, as anyone can observe, the mind is capable of comprehension rates well beyond the rate at which speech is normally articulated, and even beyond the rate at which it can accurately be produced by the vocal tract, which has a practical upper limit around three hundred words per minute. In recognition of this fact, Fletcher [30] in 1929 experimented with increased speech presentation rate. These experiments involved simple time scaling of the speech waveform by modifying the speed of a mechanical playback medium. Fletcher found that the accompanying spectral distortions, the scaling of the frequency domain by the inverse of the time domain scaler, imposed a rather narrow limit on the range over which speech so-processed is intelligible. Later work by Steinburg [31] confirmed Fletcher's basic result that intelligibility drops off rapidly from 80 per cent for time scale factors outside the interval 0.7 to 1.2. Fundamental understanding was later applied to the problem by Miller and Licklider [32], who recognized the existence of redundant information in the speech waveform, particularly during vowels and pauses. Interested primarily in taking advantage of this redundancy to facilitate time multiplexing of speech on limited bandwidth channels, they showed that periodic deletion of segments amounting to 50 per cent of the total waveform, if performed at the proper rate, would reduce intelligibility less than 10 per cent. This information was soon applied by Garvey [33] to the speech rate compression problem. Garvey investigated the possibility of concatenating the segments created by Miller and Licklider's deletions, thus producing a time signal which could be substantially shorter than the original, but whose spectral content had not been materially changed. Performed by magnetic tape cut and splice, Garvey's experiments showed better than 90 percent intelligibility for compression factors as high as 2.5, and linear decay of intelligibility to 40 per cent for a compression of 4.0. This successful method was soon automated by Fairbanks [34] and others, and remains the philosophical basis of the bulk of compression work to the present. More recent work includes the use of digitized signals which can be easily manipulated to allow pitch-synchronous splicing of segments (see references given in Chapter 1 of Portnoff [11]).

While reducing the worst of the artifacts due to arbitrary splicing, the difficulty of tracking pitch accurately, particularly where noise is present, can cause such algorithms to behave poorly.

Crude as these methods seem, they have helped define what is meant by speech compression. As Fletcher showed at the beginning, a speech compression algorithm must delineate between speech characteristics which are perceived in time and those which are perceived as having frequency significance. Furthermore, as noticed by Fairbanks and others, the preservation of waveform intelligibility requires that modifications be made over distances longer than fundamental wavelengths, but shorter than the duration over which the harmonic character of the signal can change. The division of information into what is referred to herein as temporal and harmonic information is a notion familiar to builders of vocoders, where bandwidth is greatly reduced by extracting and transmitting slowly varying harmonic information as a function of time. Hence, the vocoder is a natural tool for speech compression/expansion. In a vocoder rate change system the parameter signals produced by the analysis are compressed or expanded in time prior to synthesis. During synthesis, the implicit harmonic content is restored, unaltered. Hence, in theory, only the temporal scale is modified. Probably the highest quality result obtained in vocoder speech compression/expansion to date was reported by

Portnoff in 1978 [11]. Portnoff implemented the scheme suggested by earlier by Flanagan [1] wherein the code domain of the phase vocoder (the short-time spectral domain) was time scaled prior to synthesis. As pointed out by Portnoff, the phase vocoder is a more natural tool for compression/expansion than most other vocoder types, since it requires no tracking of formants, pitch or voiced-unvoiced information, and is, in fact, an identity system in the absence of code domain modification (such as compression, expansion, noise addition, filtering, etc.).

5.2 The Resolution Issue in Compression/Expansion

A fundamental issue in connection with vocoder compression/expansion schemes involves the fact that the analysis of a signal into its temporal and harmonic components is properly performed, not on the basis of signal modelling, but on the basis of the perceptual mechanism. As pointed out earlier, the short-time Fourier transform imposes a frequency-independent definition of what information is considered to be harmonic, and which is temporal. Hence, as Portnoff [11] writes, the builder of a compression/expansion scheme based on short-time Fourier analysis is forced

> . . . to compromise between the requirements of resolving the pitch of speech (in the frequency domain) and resolving the temporal events of the speech (in the time domain). At times, the assumption that both of these requirements can be satisfied simultaneously is a borderline assumption. [p. 140]

In an argument in favor of temporally adaptive time resolution in vocoders, Patisaul and Hammet [35] conclude that

> . . . . there is no optimum compromise in time-frequency resolution. Instead, the 'filter' nature of the hearing process, the extremes in the articulatory dynamics of speech production, the desire for the validity of the stationary model and the concept of a time-frequency cell 'matched' to the signal suggest that the shape of the resolution rectangle in vocoder spectrum analysis should be adapted to the signal. [p. 1298]

Evidence indicates that a constant bandwidth analysis is consistent with neither the human auditory system in general nor with a correct formulation of the rate compression/expansion problem in particular. For examrle, recent automatic phoneme recognition work by Searle [6] suggests that information by which various burst and stop phonemes are recognized occurs with time resolutions finer than 20 msec, and probably as fine as 5 to 10 msec. The auditory system, on the other hand, hears tones with fundamentals longer than 20 msec. Thus, the constant-Q transform, which maps signals into a two-dimensional space where time and frequency resolutions are dependent on analysis frequency, provides a more natural tool for performing independent modifications to temporal or harmonic aspects of signals. The problem, then, of defining what portions of a signal ought to be compressed or expanded in a speech rate change system is at least partially solved by requiring the time-frequency boundary to be a variable related to the ear's frequency-dependent boundary.

5.3 Constant-Q Temporal/Harmonic Compression/Expansion

The approach to rate changes taken in the work reported here utilizes a property of the constant-Q transform not shared by the short-time Fourier transform. This property, proved in Section 3.6.2 is as follows:

$$f(\alpha t) \leftrightarrow F(\omega/\alpha,\alpha t)/|\alpha| \qquad (5.1)$$

This property can be used to relate a change of scale of either the temporal or the harmonic spectral information to a change of scale of both the time domain signal and the other spectral axis. Assume, for example, the possibility of scaling the temporal axis of the constant-Q spectrum by α. This would give a new spectral function, $F'(\omega,t)$,

$$F'(\omega,t) = F(\omega,\alpha t) \qquad (5.2)$$

If the signal, $f'(t)$, resulting from substituting $F'(\omega,t)$ into 5.1 were time scaled by $1/\alpha$, the result, $F''(\omega,t)$, using the constant-Q time scaling property would be

$$F''(\omega,t) = |\alpha|F'(\alpha\omega,t/\alpha) \leftrightarrow f'(t/\alpha) \qquad (5.3)$$

This may be written in terms of $F(\omega,t)$ as,

$$F''(\omega,t) = |\alpha|F(\alpha\omega,t) \qquad (5.4)$$

Thus, as illustrated in Figure 5.1, a harmonically scaled constant-Q spectral domain is related to a temporally scaled constant-Q spectral domain by a change of the signal's time scale.

5.4 Implementation of a Constant-Q Compressor/Expander

Because of the relationship explained in the above section, independent scaling of either the temporal or the harmonic axis may be performed if one or the other is possible. Both methods will be outlined in the following sections, which review the issues involved in temporal or harmonic scaling.

5.4.1 Temporal Compression/Expansion

Modification of the scale of the temporal axis of the constant-Q spectral domain can be performed as indicated in Figure 5.2a. In this block diagram, each channel output of a continuous-time, discrete frequency analyzer is time scaled by prior to ordinary synthesis. As shown, this time scaling of discrete-time data is accomplished by resampling the data while holding the implicit sampling frequency constant. Thus, for example, if an analyzer channel output is represented as $F(\omega_k, i\Delta t(\omega_k))$, the time scaled channel data may be written as $F_\alpha(\omega_k, i\Delta t(\omega_k)) = F(\omega_k, \alpha i\Delta t(\omega_k))$. This is efficiently accomplished for any rational scalar, α, using a method such as that described by Rabiner and Crochiere [24] (see Section 4.4). However, because each channel signal is itself subject to the Fourier scaling property, this

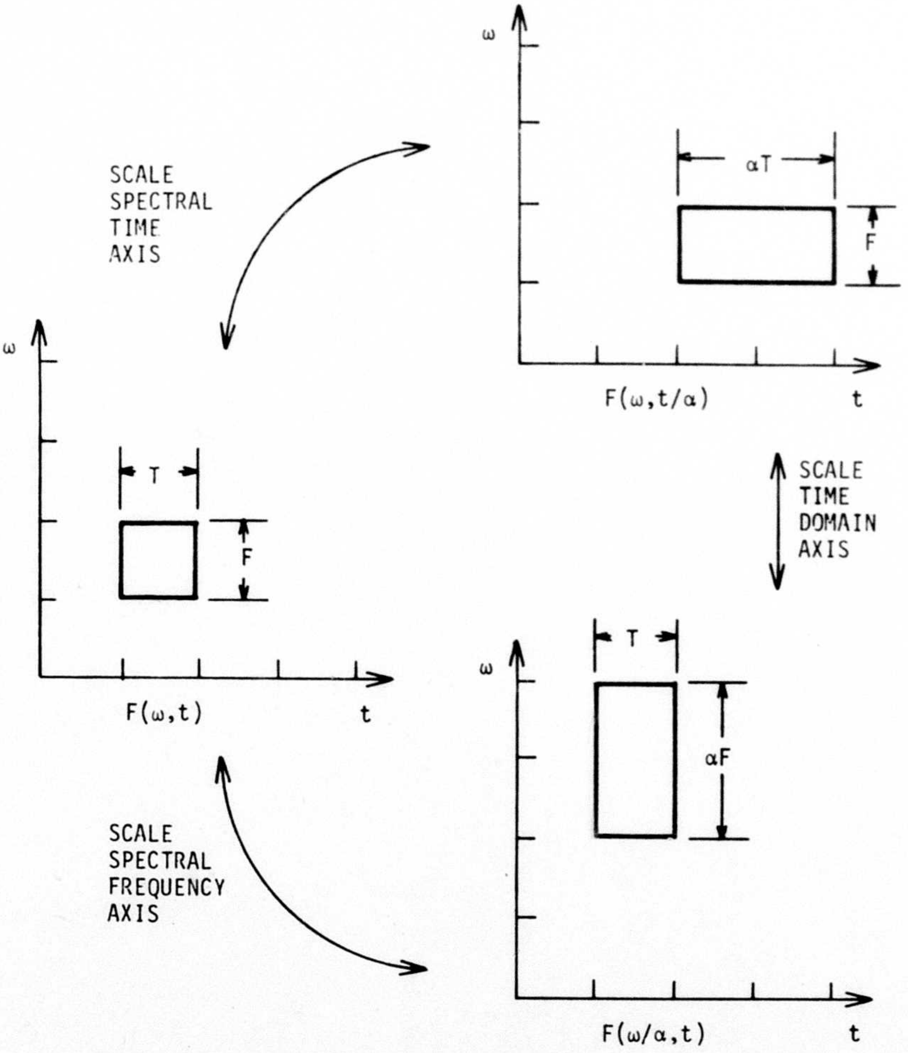


Figure 5.1. Relationship of axis scaling in the time and spectral domains.

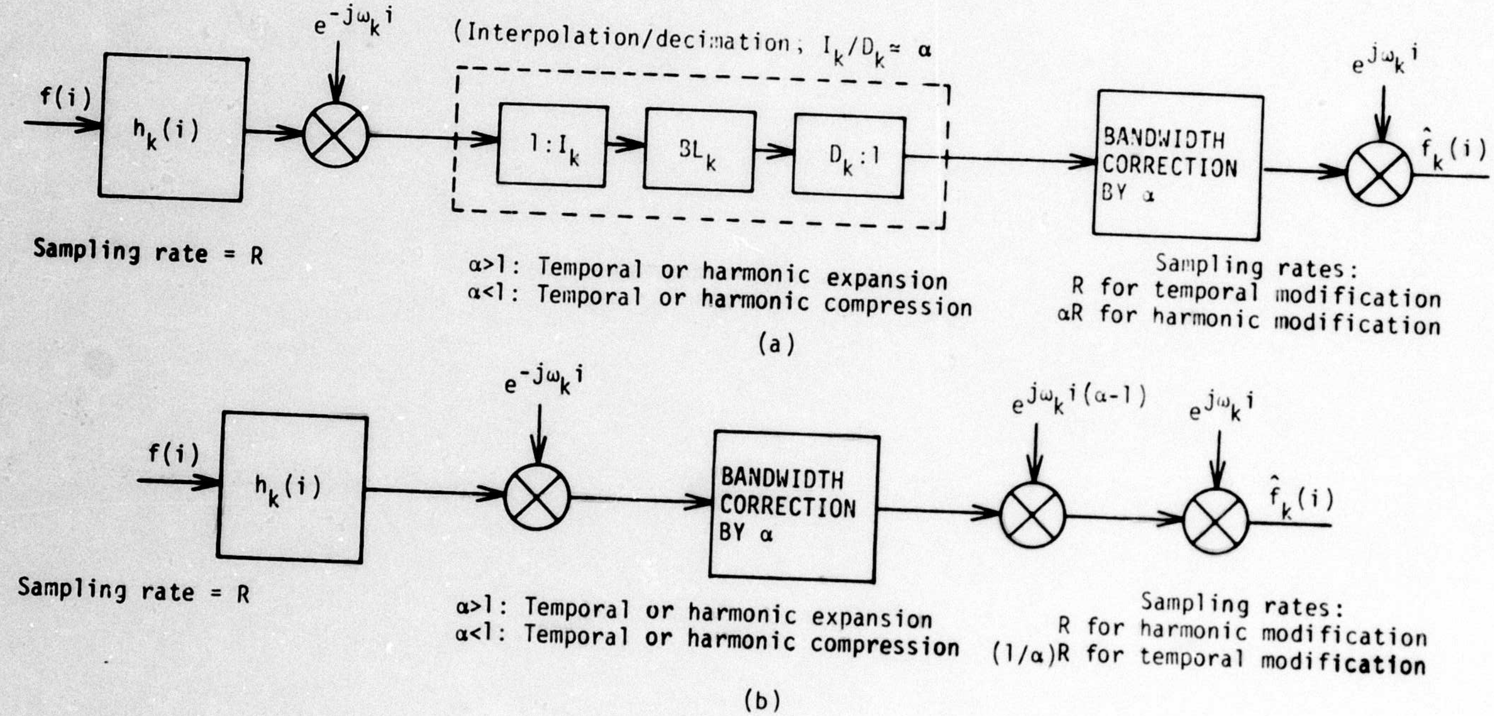

f(i)
$h_k(i)$
$e^{-j\omega_k i}$
(Interpolation/decimation; $I_k/D_k \simeq \alpha$
$1:I_k$
BL_k
$D_k:1$
BANDWIDTH CORRECTION BY α
$e^{j\omega_k i}$
$\hat{f}_k(i)$
Sampling rate = R
α>1: Temporal or harmonic expansion
α<1: Temporal or harmonic compression
Sampling rates:
R for temporal modification
αR for harmonic modification
(a)
f(i)
$h_k(i)$
$e^{-j\omega_k i}$
BANDWIDTH CORRECTION BY α
$e^{j\omega_k i(\alpha-1)}$
$e^{j\omega_k i}$
$\hat{f}_k(i)$
Sampling rate = R
α>1: Temporal or harmonic expansion
α<1: Temporal or harmonic compression
Sampling rates:
R for harmonic modification
(1/α)R for temporal modification
(b)

Figure 5.2. Independent spectral axis scaling. The upper diagram (a) shows the direct, time scaling method, while (b) shows the method wherein the frequency axis is scaled.

operation modifies each channel amplitude by $1/\alpha$, and the bandwidth of each channel by a factor of α. The correction of this unwanted effect is a fundamental issue in independent modification of the spectral axes, and will be discussed in Sections 5.4.3 and 5.5. For now, it is important to insure that the sampling period, $\Delta t(w_k)$, is adequate to represent the modified bandwidth of $F_\alpha(\omega_k,i\Delta t(\omega_k))$ without aliasing. If, as in Figure 5.2a, the analysis output is sampled at the original signal sampling rate, and if $F_\infty(\omega_k)$ is less than the signal sampling frequency, then adequate bandwidth exists.

Where α is less than unity (temporal expansion), each channel will occupy less bandwidth as the result of resampling, obviating the above condition.

5.4.2 Harmonic Compression/expansion

The above description applies to the process of independent temporal axis scaling. If the goal in scaling the temporal axis was temporal compression/expansion, the synthesis output is simply reproduced at the original sampling frequency with the necessary lowpass filter. Where the purpose was really harmonic axis scaling, the relationship of Figures 5.1b and 5.1c may be used. The temporally-scaled signal is simply time scaled by reproducing at a sampling rate equal to α times the original rate. This produces a result which occupies the original time duration, but which has been scaled along the harmonic axis (the overall operation is sometimes referred to as frequency multiplication or division).

Direct modification of the scale of the harmonic axis of the constant-Q spectral domain by is accomplished by complex modulation of the kth channel by a frequency equal to α-1 times the original channel center frequency, so that the shifted continuous-time channel signal, $F(\omega,t)$, is given by

$$F_{\alpha\omega}(\omega_k,t) = F(\omega_k,t)e^{j(\alpha-1)\omega_k t} \tag{5.5}$$

A schematic representation of a harmonic-axis scalar appears in Figure 5.2b. The harmonic shift described above does not, however, modify the bandwidth of the channel signal.

5.4.3 Bandwidth Scaling by Scaling of the Phase-derivative

The effect of this error is easily seen by the following example.

Suppose that channels of a continuous-time constant-Q filterbank appear as in Figure 5.3a as they relate to a complex input signal, x(t),

$$x(t) = e^{j\omega_0 t} \tag{5.6}$$

The analysis output for each channel can be written as

$$F(\omega_k,t) = \sigma_k e^{j\omega_0 t}e^{-j\omega_k t} \tag{5.7}$$

when

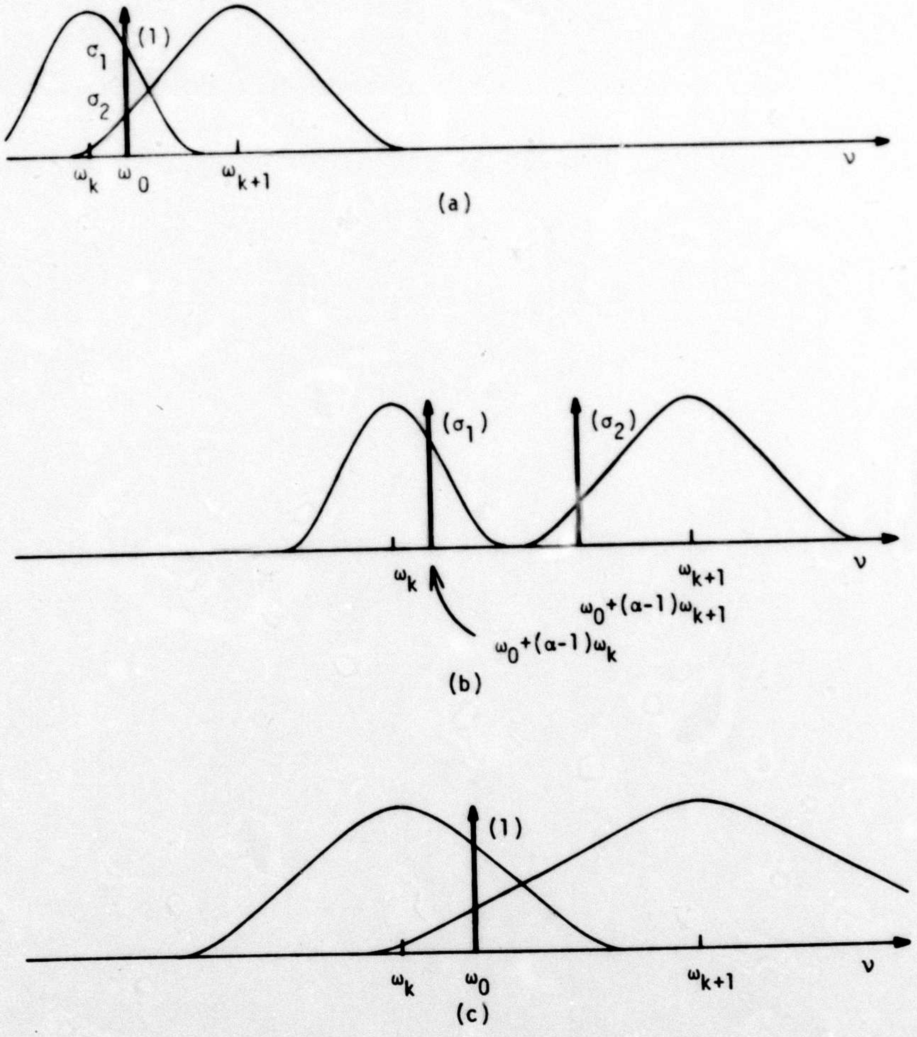


Figure 5.3. Illustration of the effect of bandwidth compensation in frequency axis scaling. The upper drawing (a) shows the original signal in relationship to the pertinent analysis filters. In (b) the channel center frequencies have been scaled, but not the effective bandwidths. Note the splitting of the signal. Finally, (c) shows the result of properly scaling both center frequencies and bandwidths.

$$\sigma_k = H((\nu-\omega_k)/\omega_k)/|\omega_k| \Big|_{\nu=\omega_0} \tag{5.8}$$

If harmonic scaling is attempted as above without bandwidth compensation, the resulting signal, $\tilde{x}_{\alpha\omega}(t)$, is given by

$$\tilde{x}_{\alpha\omega}(t) = \sum_k F(\omega_k,t)e^{j(\alpha-1)\omega_k t}e^{j\omega_k t} \tag{5.9}$$

$$\tilde{x}_{\alpha\omega}(t) = \sum_k \sigma_k e^{j(\omega_0+(\alpha-1)\omega_k)t} \tag{5.10}$$

As illustrated in Figure 5.3b, this result is not simply a shifted complex sinusoid, but a sum of scaled, unequally shifted sinusoids. In this example, the problem is solved by scaling the phase of the analysis output by α prior to shifting and synthesis. Then

$$x_{\alpha\omega}(t) = \sum_k \sigma_k e^{j(\alpha(\omega_0-\omega_k)t}e^{j(\alpha-1)\omega_k t}e^{j\omega_k t} \tag{5.11}$$

$$x_{\alpha\omega}(t) = e^{j\alpha\omega_0 t}\sum_k \sigma_k \tag{5.12}$$

Since we require the magnitude of the composite filterbank to be unity, we may write

$$\sum_k \sigma_k \simeq 1 \tag{5.13}$$

and

$$x_{\alpha\omega}(t) \simeq e^{j\alpha\omega_0 t} \tag{5.14}$$

As shown in Figure 5.3c, the result is the original complex sinusoid with scaled center frequency.

The operation by which the bandwidth was scaled in the above example, scaling of the constant-Q spectral phase, is equivalent to that suggested by Flanagan [1] in connection with a compression/expansion system based on the phase vocoder. The effect of scaling the constant-Q spectral phase is understood by recognizing that it equivalently scales the phase derivative (the derivative being a linear operator.) The phase derivative is an indication of instantaneous frequency, so that scaling this quantity could be thought of as scaling the bandwidth (Appendix D). A single channel of a constant-Q harmonic scaler, including bandwidth expansion, appears in Figure 5.2b.

It should be noted as above that, in a discrete-time implementation, adequate bandwidth should be allowed in each channel signal when harmonic expansion is performed. For instance, if each channel analysis output were minimally sampled, bandwidth expansion by any factor would cause the channel signal to be undersampled (and hence aliased) by that factor. Following bandwidth compensation, if harmonic scale modification was the goal, the resulting synthesized signal is reproduced at the original sampling frequency using proper anti-imaging filters. If indirect temporal scaling was the desired end, the reproduce sampling frequency and the anti-imaging filter cutoff should be scaled by α.

5.5 Bandwidth Scaling Issues

Although the scaling of the constant-Q spectral phase time-derivative, proposed above as a means of modifying channel bandwidth, was used in this research with good results, another approach suggests that the scaling of the phase derivative may be only an approximate solution to the problem of bandwidth scaling. This different approach will be described in the present section.

The problem of scaling the bandwidth of a complex analysis output channel can be best solved if (1) a model describing such signals in a general way exists and if (2) a relation exists which measures bandwidth in terms of the model's parameters. Such a relationship has been described by Kahn and Thomas [36] for amplitude and and angle modulated (AAM) signals of the form

$$x(t) = M(t)e^{j\theta(t)} \tag{5.15}$$

In this model, M(t) is an amplitude-modulating function equivalent to the constant-Q spectral magnitude, and $\theta(t)$ is a phase modulating function equivalent to the constant-Q spectral phase. Kahn and Thomas point out that, in general, the modulating functions may have infinite bandwidth, but that in practice, most spectral information is concentrated within a finite band. They propose, as a useful measure of this bandwidth, a second moment measure of the spread of the power spectral density. If by $S_x(t,\omega)$ we represent the power spectral density function of x(t), and if $R_x(t,\tau)$ equals the autocorrelation function, then

$$S_x(t,\omega) = \int R_x(t,\tau)e^{-j\omega\tau}\, d\tau \tag{5.16}$$

and

$$R_x(t,\tau) = E\{x(t)x^*(t-\tau)\} \tag{5.17}$$

where E{} denotes mathematical expectation. The instantaneous bandwidth, $\Omega_x(t)$, can be defined by the normalized second moment as follows:

$$\Omega_x(t) = \{\int \omega^2|S_x(t,\omega)|^2\, d\omega\}^{\frac{1}{2}}/\{\int |S_x(t,\omega)|^2\, d\omega\}^{\frac{1}{2}} \tag{5.18}$$

This may be rewritten, using Parseval's theorem and the differentiation property of the Fourier integral transform [14], as

$$\Omega_x(t) = \left\{\frac{-\delta^2 R_x(t,\tau)}{\delta t^2}\right\}^{\frac{1}{2}} /\{R_x(t,\tau)\}^{\frac{1}{2}} \Bigg|_{\tau=0} \tag{5.19}$$

$$\Omega_x(t) = |E\{|\dot{x}(t)|^2\}/E\{|x(t)|^2\}|^{\frac{1}{2}} \tag{5.20}$$

Substituting x(t) from 5.15, and assuming M(t) and θ(t) to be real-valued and differentiable, and $E\{M(t)\dot{M}(t)\dot{\theta}(t)\}=0$,

$$\Omega_x(t) = \left\{\frac{E\{\dot{M}^2(t)+M^2(t)\dot{\theta}^2(t)\}}{E\{M^2(t)\}}\right\}^{\frac{1}{2}} \tag{5.21}$$

This can be written, in light of 5.19, as

$$\Omega_x^2(t) = \Omega_M^2(t) + \frac{E\{M^2(t)\dot{\theta}^2(t)\}}{E\{M^2(t)\}} \qquad (5.22)$$

An interpretation of this result is simply that the total bandwidth of an AAM signal has components which are due to (1) the bandwidth of the amplitude modulating signal, and to (2) the phase modulating signal. In the case where the amplitude modulating signal is a very slowly varying function, 5.22 depends linearly on the phase modulating function, and the approximation presented in Section 5.4 is exact. If, however, the amplitude modulation portion of the bandwidth is a significant, but not dominant, portion of the total bandwidth, simple phase scaling will lead to inaccurately-scaled bands. This error, in cases where bandwidth is being expanded, leads to misplaced spectral data, and possible spectral holes, giving rise in extreme cases to an effect similar to comb filtering. To determine a corrected factor by which the phase derivative should be scaled, assume that a correct bandwidth-expanded signal, C_α, is given by,

$$C_\alpha(t) = M_\alpha(t)\cos\{\omega_k t+\theta_\alpha(t)\} \qquad (5.23)$$

and that,

$$M_\alpha(t) = \alpha_1 M(t) \qquad (5.24)$$

$$\theta_{\alpha}(t) = \alpha_2\theta(t) \qquad (5.25)$$

where α_1 and α_2 must be real. Then,

$$\alpha^2\Omega^2_{C_\alpha} = \frac{E\{\dot{M}^2(t)\}+\alpha_2E\{M^2(t)\dot{\theta}^2(t)\}}{E\{M^2(t)\}} \qquad (5.26)$$

Note that α_1 has no effect. Substituting the equation 5.22 expression for Ω_{C_α} into 5.26, the value of α_2 can be determined.

$$\alpha_2 = \alpha\left[1+(1-\alpha^{-2})E\{\dot{M}^2(t)\}/E\{M^2(t)\dot{\theta}^2(t)\}\right]^{\frac{1}{2}} \qquad (5.27)$$

This equation implies a conditional relationship between α and α_2 . When expanding bands (α>1), the number by which the phase derivative must be scaled to scale the bandwidth by α is greater than α. When contracting bands (α<1), the opposite is true. It should be noted that when the amplitude modulation contribution, $E\{\dot{M}^2(t)\}/E\{M^2(t)\}$, dominates the total bandwidth, 5.27 may become imaginary. This effect corresponds to a lower limit on the total bandwidth reduction available using the assumptions of equations 5.24 and 5.25.

5.6 Results

The procedure outlined in this chapter for the independent time or frequency scaling of signals was found to produce good quality results over a range of scaling parameters. Rate compression was achieved for factors up to four, at which point intelligibility was degraded primarily because of the inability of the mind to assimilate rapidly enough. It was also noted that phonemes which occur naturally over very short intervals tend to disappear at high compression ratios. This phenomenon represents a fundamental limit to the uniform definition of compression adopted here, which can be circumvented only by selectively compressing features in a non-uniform manner. Such an investigation is beyond the scope of this work.

Expansion experiments were successfully performed for factors as low as one-third. The speech expansion experiments revealed two fundamental difficulties in constant-Q expansion.

First, for values for Q thought to be comparable to the selectivity of the human auditory system (see Section 1.2), the constant-Q transform was found to have time-resolution at high frequencies which was too fine. As shown in Figure 5.4a, the magnitude of high frequency formant information appears to be modulated pitch synchronously. The result is that when the time scale of the high bands is stretched, and the bands are remodulated, a false modulation of these bands occurs at a frequency equal to the old pitch scaled by the expansion factor. In expansion, this results in subjective effects best described as granularity or roughness of the voice. This effect is reduced or eliminated by increasing the selectivity of the analysis, thus reducing the fineness of

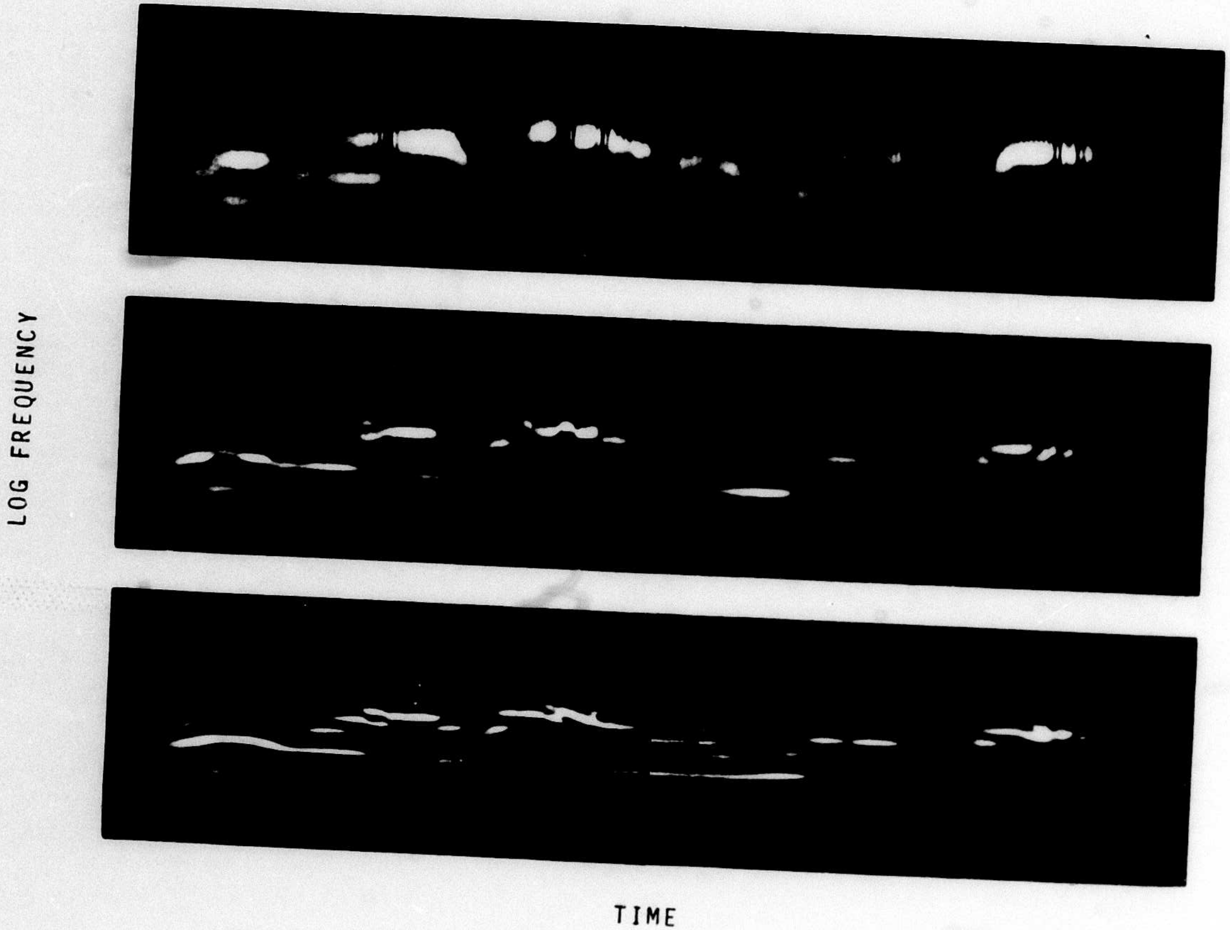


Figure 5.4. Comparison of spectrograms of various selectivities. The selectivities for the various spectrograms are respectively, Q=3.0, Q=8.0, and Q=12.0.

the transform's time resolution. Spectrograms with reduced time resolution (increased Q) are shown in Figures 5.4b and 5.4c. Temporally expanded synthesis was performed using Q=19. This adjustment corrected the granularity problem without introducing any undesirable side effects.

The second difficulty which was made evident by the expansion experiments was the presence of artifacts due to spectral phase unwrapping errors. Such errors may be eliminated for compressions by even integer factors when unwrapping methods are used which modify the wrapped phase only by integer multiples of π . In such cases any error becomes effectively zero (a multiple of 2π) during the expansion of the individual bands. In speech which has not been expanded by even integer factors, errors of this variety, because of their random nature, produce a low-level "gurbling" sound. This effect results because the phase shifts produce random constructive or destructive interference among overlapping channels.

Because of the difficulty in properly characterizing the spectral phase function, tests by which the unwrappers mentioned in Section 4.6 could be evaluated were difficult to design. The above-mentioned property of expansion by even integer factors provided one good test. In such a test, wrapped and unwrapped spectral phase functions were both temporally expanded, synthesized, and compared. Error other than integer π errors appeared as differences between the two syntheses. Comparison of the effect of expansion of a given unwrapper output for even integer factors and non-integer factors provided a means of isolating the other category of unwrapper errors -- namely errors of π.

CHAPTER 6

CONCLUSION

6.1 Summary

The potential ability of a constant percentage bandwidth transform to model human auditory analysis of sound has been discussed, as have previous efforts to simulate such analysis. A variation of the Gambardella "multiple filter analyzer integral," referred to herein as the constant-Q transform (CQT) has been shown to be capable of constant-Q analysis of signals.

The major contribution of this work consists of the establishment of the CQT as an understood, theoretically sound tool for audio signal processing. To this end, a synthesis transform has been proposed which performs a reverse mapping from the constant-Q spectral domain to the time domain, and which, in the absence of spectral modifications, forms an identity system when cascaded with the analysis transformation. Existence conditions, and a proof of the validity of this synthesis transformation have been given. Both the forward and the reverse transformations have been compared, using a filterbank analogy to the more familier short-time Fourier transform, and similarities and important differences have been pointed out.

The effect of spectral modification has been established for both the constant-Q transform and for a generalized form of the short-time Fourier transform. The dependence of the effect of spectral modification upon the analysis window has been established. Several useful transform properties have been listed and proved.

Principles governing the sampling of the CQT without loss of information have been formalized, and explicit expressions have been derived which relate the characteristics of the analysis window with the CQT selectivity and with a minimal CQT spectral domain sampling pattern. An alternative logarithmically warped form of the CQT synthesis was also established which enables minimal sampling of the spectral domain. The minimum overall sampling rate for the constant-Q spectrum was shown to be equivalent to the short-time Fourier transform and to the time-domain sampling rates. Two practical analysis-synthesis algorithms were discussed -- one which achieves nearly minimal spectral sampling, and another which achieves a simpler implementation at the expense of over sampling in time.

The nature and computation of the constant-Q spectral magnitude and phase functions has been investigated. A method for unwrapping phase in the difficult broad band high channels was proposed which attempts to perform band limiting correction on wrapped phase using a phase derivative estimate computed directly from the real and imaginary parts. The algorithm was shown empirically to perform better than two other more common methods.

The first high-resolution constant-Q spectrograms of speech have been produced.

Finally, the usefulness of the CQT in actual signal processing has been demonstrated by application to the perception-related problem of rate modification of speech. Good quality modification was achieved for rates between 1/2 and 4. Limitations, some inherent to the notion of rate changing, some resulting from the nature of a CQT implementation, and some computational, were explained.

6.2 Further Research and Suggested Applications

The CQT has been established as a well-defined tool for audio processing. Its potential usefulness, however, appears to be limited primarily by lack of a fast computational algorithm analogous to the FFT. Processing times on a dedicated PDP-10 processor are currently on the order of 10^3 times real time. The development of fast algorithms for computing and processing minimally-sampled CQT spectral data is an area which merits future investigation.

Other general areas not related to implementational and computational issues are naturally suggested by the analogy that exists between the CQT and the peripheral auditory system. Noise suppression using two-dimensional spectral subtraction or thresholding would be able to track more closely in time rapidly changing high frequency information, while better resolving in frequency the low frequency portions of a signal. This frequency-specific resolution might reduce effects related to leakage of noise in areas where the ear is capable of fine resolution.

Acoustic enhancement experiments analogous to the well-known visual enhancement procedures seem promising. Constant bandwidth experiments, performed by Callahan using two-dimensional techniques suggest the potential of such experiments if performed in the constant-Q spectral domain.

Studies by Searle [6] indicate that a transform which more closely emulates the analysis performed by the human ear could be advantageous in the automated recognition of speech. Information essential to the recognition of stops and bursts seems to occur with resolution finer than conventional analysis provides. As constant-Q algorithm speed improves, a system based on two-dimensional constant-Q recognition merits investigation.

Finally, the similarities between the distribution of information in the constant-Q spectral domain and auditory analysis suggest uniform quantization of a minimally-sampled constant-Q spectrum as a means of reducing overall bandwidth at minimal perceptual expense.

APPENDIX A

RESOLUTION PROPERTIES OF WINDOWS

The scaling property of the Fourier integral transform indicates a reciprocal relationship between the scale of events measured in the time and frequency domains. Specifically, if h(t) is a time function defined over all t, and if the integral,

$$H(\omega) = \int h(t)e^{-j\omega t}\, dt \qquad (A.1)$$

exists, then it is true that

$$H(\omega/\alpha)/|\alpha| = \int h(\alpha t)e^{-j\omega t}\, dt \qquad (A.2)$$

If we define some characteristic time length, T, and a characteristic frequency length, F, then the above relationship guarantees that the product, β, of T and F is a constant.

$$\beta = TF \qquad (A.3)$$

The value of β is wholly dependent on the function, f(t), and on the definitions of T and F. This property provides a simple way of relating the time and frequency resolution of an analysis window. We adopt the convention that resolutions in either domain will be measured as the width of the principal interval centered on zero during which the function is attenuated less than σ decibels from its maximum value. In the time domain, for common windows such as the Fourier, Hann, Hamming, Blackman and Bartlet windows, we pick $\sigma = \infty$. Thus, T_∞ is the total non-zero length of a window. In the frequency domain, two measures are useful, F_∞ and the so-called 3 decibel bandwidth, F_3. From these we may define and compute values for the analysis window resolution products,

$$\beta_\infty = T_\infty F_\infty \qquad \text{(A.4)}$$

$$\beta_3 = T_\infty F_3 \qquad \text{(A.5)}$$

Table A.1 lists values of β_∞ and β_3 for the windows mentioned above. Values for β_∞ are exact and can be arrived at analytically; values for β_3 are determined numerically.

Table A.1

RESOLUTION PRODUCTS FOR SOME COMMON WINDOWS

WINDOW	FUNCTION	β	β_2
Bartlet	$h(t)=[1-\|2t\|]p(t)$	4.0	1.273560
Blackman	$h(t)=[0.42+0.5\cos(2\pi t)-0.08\cos(4\ t)]$	6.0	1.640929
Fourier	$h(t)=p(t)$	2.0	0.884487
Hamming	$h(t)=[0.54+0.46\cos(2\pi t)]p(t)$	4.0	1.300817
Hann	$h(t)=[0.5+0.5\cos(2\pi t)]p(t)$	4.0	1.438205

Note that $p(t)$ is the pulse which is zero except in the interval, $(-1/2,1/2)$.

APPENDIX B

SUPPLEMENTAL MATERIAL RELATIVE TO GENERALIZED SHORT-TIME FOURIER SYNTHESIS

B.1 Validity of Generalized Short-time Fourier Synthesis

The continuous short-time Fourier transform, $F(\omega,t)$, of a function, $f(t)$, is given by

$$F(\omega,t) = \int f(\tau)h(t-\tau)e^{-j\omega\tau}\, d\tau \tag{B.1}$$

Allen and Rabiner [15] have described two commonly-understood methods for synthesis of $f(t)$ from $F(\omega,t)$. The two syntheses, the filterbank summation (FBS) method and the overlap-add (OLA) method are given in their continuous forms by equations B.2 and B.3, respectively.

$$f(t) = \int F(\omega,t)e^{j\omega t}\, d\omega/2\pi h(0) \tag{B.2}$$

$$f(t) = \iint F(\omega,\tau)e^{j\omega t}\, d\omega\, d\tau/2\pi \tag{B.3}$$

The analysis maps signals of the form, $f(t)$, from the line into a subclass of the plane in such a way that, in the absence of spectral domain modifications, the reverse mapping can be made with perfect fidelity using either B.2 or B.3. The set of spectral domain signals of the form, $F(\omega,t)$, which is reachable via B.1 is restricted in resolution by the analysis window, $h(t)$. Signals which are not so-constrained are, nevertheless, also reverse mappable onto the line by members of a set of functions, called retracts, of which B.2 and B.3 are examples. The way in which this reverse mapping occurs is determined by the form of the retract, and is of importance when considering the effects of spectral domain modifications which violate time or frequency resolution constraints placed on spectral domain signals by h(t). The OLA and FBS synthesis do not exhaust the possible forms of a more general class of retracts useable for short-time Fourier synthesis. Rather, they are special cases of the general retract,

$$f(t) = \frac{1}{2\pi \langle g,h \rangle} \iint F(\omega,\tau) g(t-\tau) e^{j\omega t} \, d\omega \, d\tau \qquad (B.4)$$

where <g,h> is the inner product,

$$\langle g,h \rangle = \int g(t) h(t) \, dt \qquad (B.5)$$

The FBS and OLA synthesis are easily seen to be special forms of B.4. In particular, the FBS synthesis is obtained given the condition,

$$g(t) = \delta(t) \qquad (B.6)$$

where $\delta(t)$ is the Dirac delta function. Similarly, the OLA synthesis is obtained if

$$g(t) = 1 \qquad (B.7)$$

and if the area of the analysis window is unity.

We now show that B.1 and B.4 form an

analysis-synthesis identity by substituting B.1 into B.4 (renaming the result).

$$\tilde{f}(t) = \frac{1}{2\pi\langle g,h\rangle} \iiint f(\xi)h(\tau-\xi)e^{-j\omega\xi}\, d\xi\; g(t-\iota)e^{j\omega t} d\omega\; d\tau \quad (B.8)$$

Modifying the order of integration, the complex exponentials may be combined and isolated,

$$\tilde{f}(t) = \frac{1}{\langle g,h\rangle} \int f(\xi) \int g(t-\tau)h(t-\xi)\frac{1}{2\pi} \int e^{j\omega(t-\xi)} d\omega d\tau d\xi \quad (B.9)$$

and then integrated,

$$\tilde{f}(t) = \frac{1}{\langle g,h\rangle} \int f(\xi)\delta(t-\xi) \int g(t-\tau)h(\tau-\xi)\; d\tau\; d\xi \qquad (B.10)$$

With the change of variables, $\mu = t-\tau$,

$$\tilde{f}(t) = \frac{1}{\langle g,h\rangle} \int f(\xi)\delta(t-\xi) \int g(\mu)h(t-\xi-\mu)\; d\mu\; d\xi \qquad (B.11)$$

If we then define

$$p(x) = \int g(\mu)h(x-\mu)\; d\mu \qquad (B.12)$$

then B.10 may be simplified to

$$\tilde{f}(t) = \frac{1}{\langle g,h\rangle} \int f(\xi)\delta(t-\xi)p(t-\xi)\; d\xi \qquad (B.13)$$

$$\tilde{f}(t) = \frac{1}{\langle g,h\rangle} f(t)p(0) \qquad (B.14)$$

Finally, we recognize from B.12 that p(0) is just the inner product, $\langle g,h\rangle$. Hence $\tilde{f}(t) = f(t)$, and the validity of B.4 as a retract of B.1 is shown.

B.2 Intuitive Description of Short-time Synthesis Issues

The meaning of B.4 is better understood by performing the indicated integration with respect to ω. This gives,

$$f(t) = \frac{1}{<g,h>}\int F_t(t,\tau)g(t-\tau)\ d\tau \qquad (B.15)$$

where$F_t(t,\tau)$ denotes the one-dimensional inverse Fourier integral transform of $F(\omega,\tau)$ with respect to its first (frequency) parameter. An expression for the function, $F_t(t,\tau)$, may also be derived from B.1 by performing an inverse Fourier integral transform along the ω-axis. This yields,

$$F_t(t-\tau) = f(t)h(t-\tau) \qquad (B.16)$$

Note the change of variables. The time variable of the short-time Fourier transform has been renamed τ . The function, $F_t(t,\tau)$ is illustrated in Figure B.]a for a unit pulse input,

$$f(t) = \begin{cases} 1 & 1\leq t\leq 2 \\ 0 & \text{otherwise} \end{cases} \qquad (B.17)$$

and h(t) is is a Hann window.

Two methods of synthesis are obvious from the figure. The first, corresponding to the FBS synthesis of B.2 is achieved by evaluating $F_t(t,\tau)$ along the line $t=\tau$ (i.e. $g(t)=\delta(t)$ in B.15.) As explained in Section 2.4, modifications made along the spectral time (τ) axis are seen to "take effect" instantaneously in time -- no time-resolution limiting occurs. One could, for instance, time-limit the short-time spectrum at $\tau=1/2$ and find that

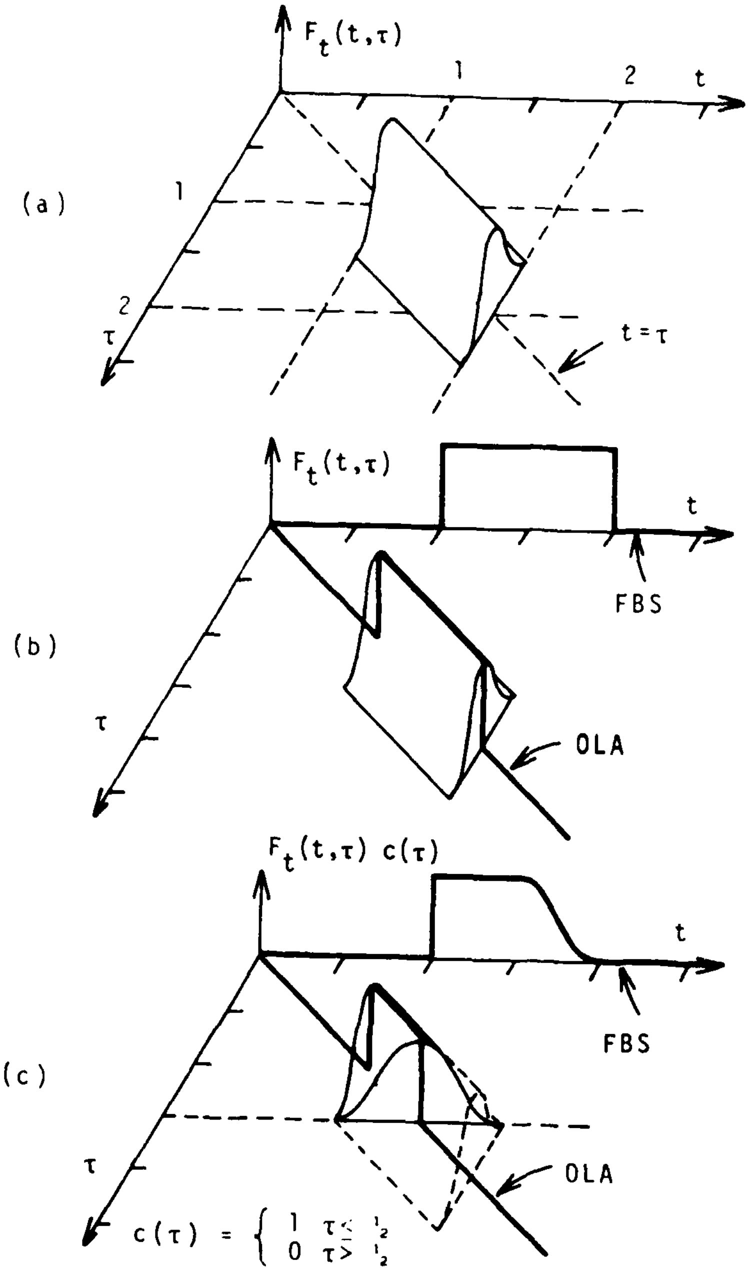


Figure B.1. Effect of frequency-independent short-time spectral modifications.

the FBS synthesis had been similarly truncated (Figure B.1b).

The other method of synthesis, the CLA synthesis of B.3, is achieved by integrating (or projecting) $F_t(t,\tau)$ along the τ axis. This corresponds to $g(t)=1$ in B.15. Note here that changes to the short-time spectrum are limited in their time resolution by the shape of the analysis window. For instance, an attempt to time-limit the spectrum as above would give the result shown in Figure B.1c.

In generalized synthesis, we pick $g(t)$ somewhere between the extremes of $\delta(t)$ and 1. The result of the convolution of B.15 then causes effective changes along the time (τ) axis of the short-time spectrum to be resolution-limited by the synthesis window, $g(t)$.

Suppose now that the spectral modifications for the extreme cases above occurred, not as a function of time (along the τ axis), but in frequency (along the transformed t axis). In this case, we might expect the modified $F_t(t,\tau)$ to appear as in Figure B.2a. Evaluating along the diagonal, as in FBS synthesis, the edges of our synthesized pulse cannot "ring" beyond the interval, (1/2,3/2) allowed by the analysis window (see Figure B.2b). Hence the attempted modification is time-limited to the dimension of the analysis window.

However, if we evaluate by integration along τ, as shown in Figure B.2c, no time-limiting occurs, and the

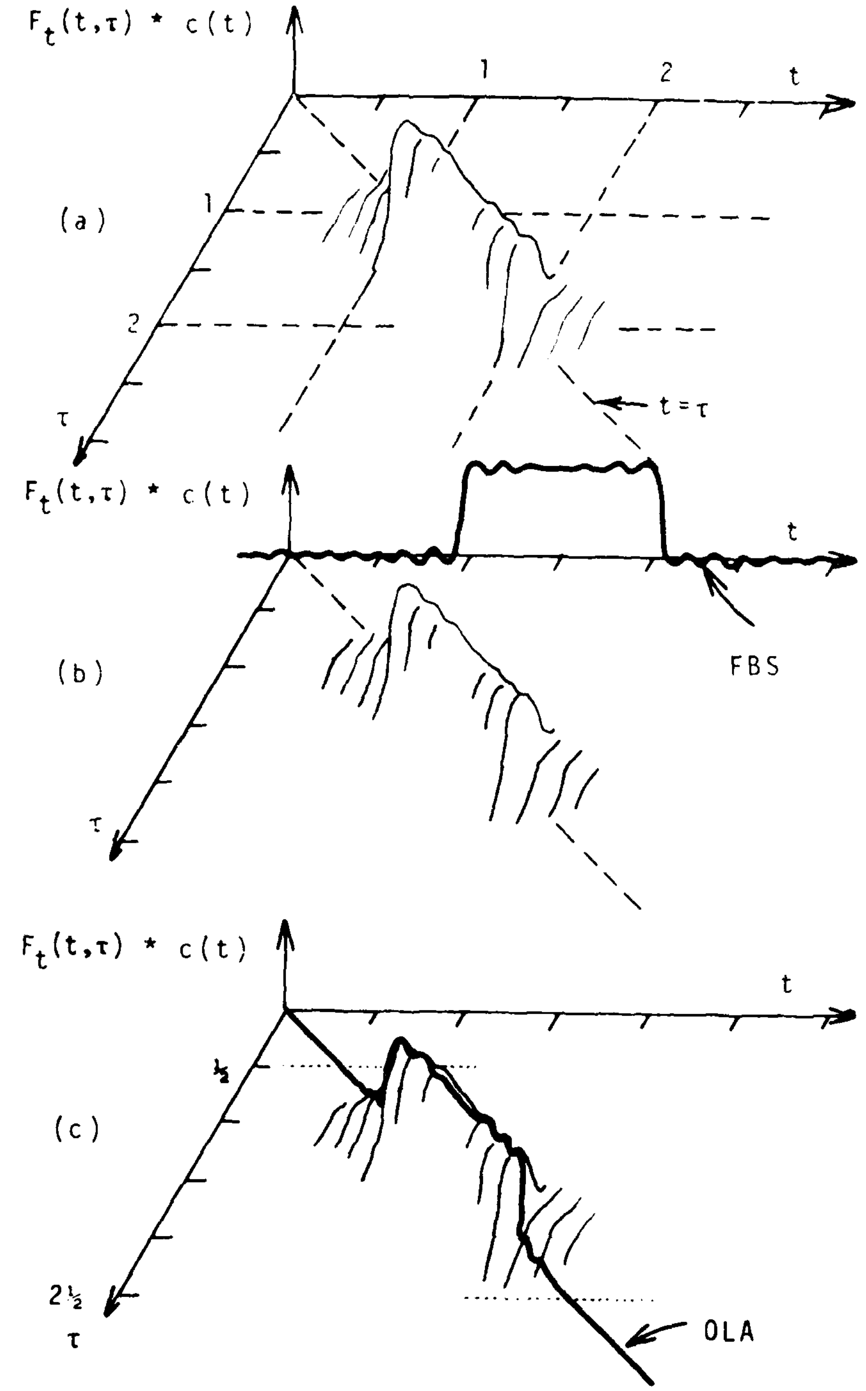


Figure B.2. Effect of time-independent short-time spectral modifications.

modification is allowed with perfect fidelity.

Again, by picking the synthesis window, g(t) somewhere between $\delta(t)$ and 1, the extent of time-limiting of frequency axis modifications is controllable.

APPENDIX C

CONDITIONS FOR THE CONVERGENCE OF CONSTANT-Q SYNTHESIS

Because analysis window functions have been defined to to have finite non-zero extent, their Fourier transforms do not. Hence, the existence of the synthesis integral,

$$\Phi(\nu) = \int \phi(\omega,\nu)\ d\nu \qquad (C.1)$$

where

$$\phi(\omega,\nu) = H((\nu-\omega)/\omega)/|\omega| \qquad (C.2)$$

(see 3.8) cannot be guaranteed on the basis of the integrand's having non-zero value over a finite interval. To establish the existence of C.1, it will be necessary to determine restrictions on h(t) which guarantee the integral's existence. This can be accomplished for a fairly general class of windows using the following assumptions. First, assume that if the Fourier integral transform of h(t) is H(x), H(x) can be bounded above for x exclusive of the interval, (-2,0), by $\sigma_1/|x+1|$ for some finite constant, σ_1. This is the case for many windows having finite non-zero time extent, since their transforms can be expressed as a sum of shifted, scaled sin(x)/x functions. It is true in particular that the windows mentioned in Table A.1 decay as 1/x. Second, assume that $H(x)$ can be bounded by $\sigma_2|x+1|$ in the interval (-2,0) for some finite constant, σ_2. This restriction is realized for the above-mentioned windows only if we restrict available values of Q to a discrete set. The Hann window, as shown in Figure C.1, satisfies this criterion when

$$\int h(t)\sin(t)\ dt = 0 \tag{C.3}$$

for n=2,3,4,.... In terms of Q,

$$Q = \frac{\beta_\infty}{\beta_3} \frac{\omega_k}{F_\infty(\omega_k)} = \frac{n\beta_\infty}{4\beta_3} \tag{C.4}$$

$$Q = .6953114n \qquad n = 2,3,4,... \tag{C.5}$$

Given the above and a σ which is at least as large as the maximum of σ_1 and σ_2, we construct a function, $B(x)$, which bounds $H(x)$ as shown in Figure C.2a. We are now prepared to examine the integrability of C.1. Notice in Figure C.2b the effect of the change of variables, $x=(\nu-\omega)/\omega$. Recognizing that $B(x) \geq H(x)$ implies $B((\nu-\omega)/\omega) \geq H((\nu-\omega)/\omega)$, we conclude that the integrability of $B((\nu-\omega)/\omega)/|\omega|$ will imply the integrability of $H((\nu-\omega)/\omega)/|\omega|$. To establish the latter implication, we integrate the bounding function piecewise as indicated in Figure C.2b. This we write as

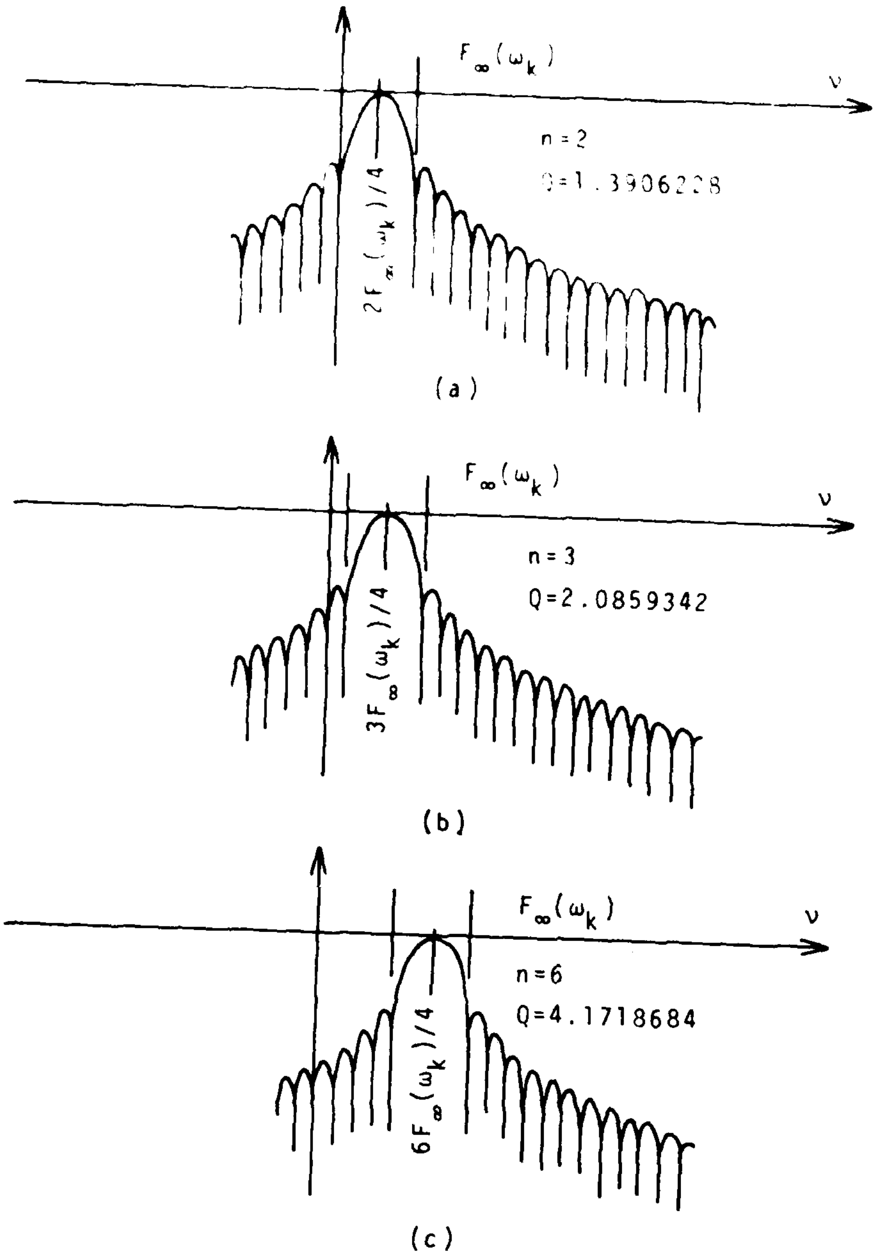


Figure C.1. Descretization of the set of allowed analysis selectivities. The three graphs show the arrangements giving rise to the n=2, n=3, and n=4 values of Q allowed for the Hann analysis window.

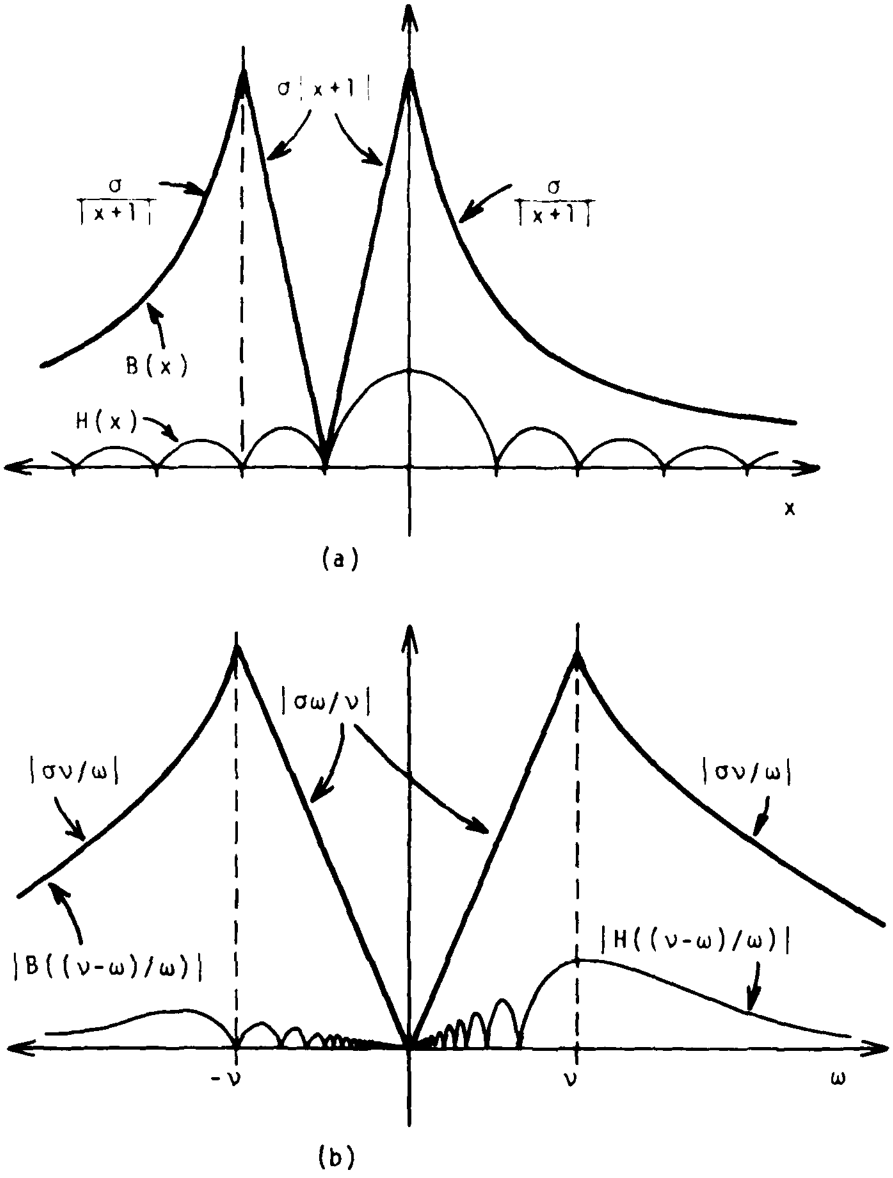


Figure C.2. A bounding function which guarantees the existence of the constant-Q synthesis integral for a Hann analysis window. The bounding function is indicated by the heavy line in both (a) and (b).

$$\Phi(\nu) \leq \int B((\nu-\omega)/\omega)/|\omega|\ d\omega \qquad (C.6)$$

which simplifies to

$$\Phi(\nu) \leq 4\sigma \qquad (C.7)$$

Thus, our rather loose bound has established the existence of $\Phi(\nu)$ for window functions which satisfy the criterion,

$$|H(x)| \leq B(x) \qquad x \in (-\infty,\infty) \qquad (C.8)$$

where

$$B(x) = \begin{cases} -\sigma/(x+1) & x<-2 \\ -\sigma(x+1) & -2\leq x<-1 \\ \sigma(x+1) & -1\leq x<0 \\ \sigma/(x+1) & 0\leq x \end{cases} \qquad (C.9)$$

for some finite σ.

APPENDIX D

THE PHASE DERIVATIVE AS A MEASURE OF INSTANTANEOUS FREQUENCY AND BANDWIDTH

A general complex signal may be represented as,

$$x(t) = a(t)e^{j(\omega t+\phi(t))} \tag{D.1}$$

where a(t) and $\phi(t)$ may respectively be thought of as the amplitude and phase modulating functions. The quantity,

$$\Xi(t) = \omega t+\phi(t) \tag{D.2}$$

provides information about instantaneous frequency and overall signal bandwidth. This line of thought is understood by observing the components of $\phi(t)$ as in Figure D.1. As seen in this figure, if $\phi(t)$ were zero, the frequency of the complex sinusoid, $e^{j\omega t}$, would equal the the slope, ω, of the line, ωt. For $\phi(t)$ non-zero, the slope is not a constant, and thus the "frequency" of the complex sinusoid, $e^{j(\omega t+\phi(t))}$, must be measured instantaneously as,

$$\frac{d}{dt}\left[\omega t+\phi(t)\right] = \omega+\dot{\phi}(t) \tag{D.3}$$

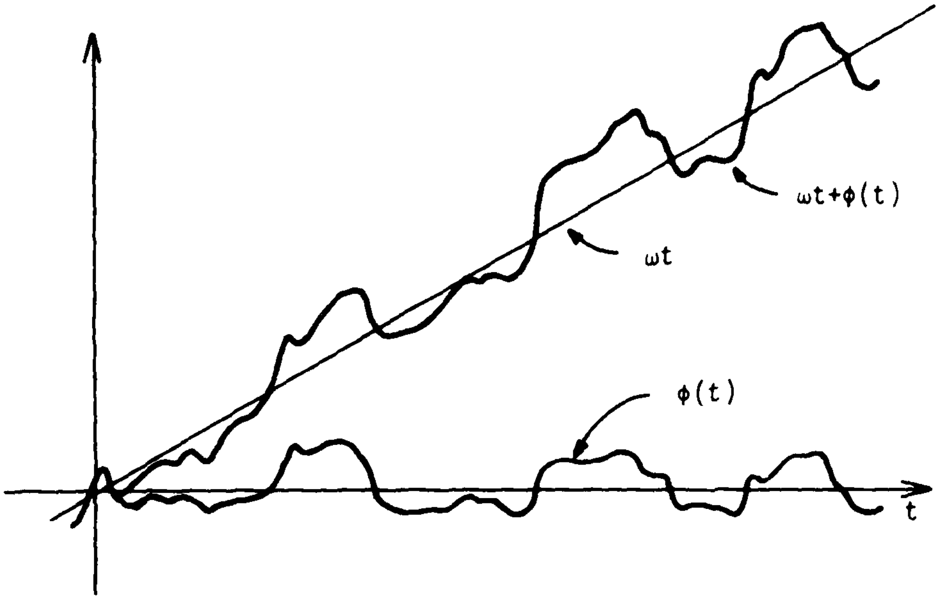


Figure D.1. Phase and center frequency.

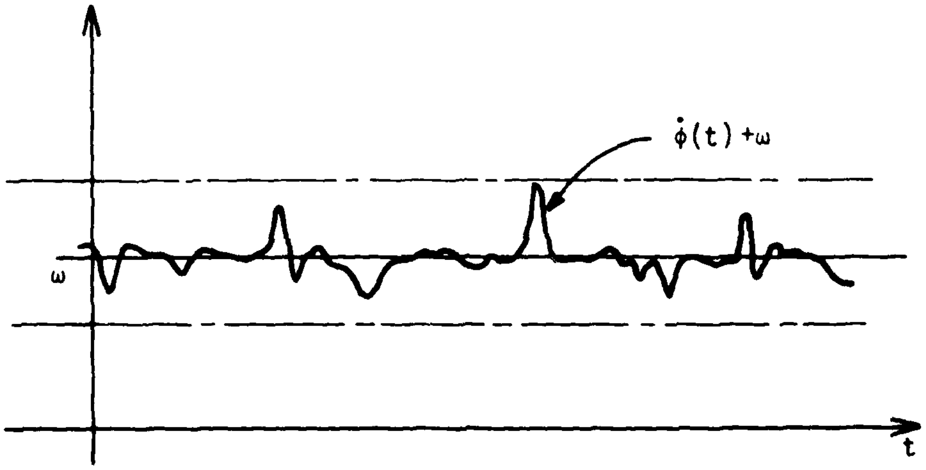


Figure D.2. Phase-derivative and bandwidth.

Thus, (t) can be thought of as an instantaneous perturbation of the center frequency, ω -- an instantaneous frequency.

Plotting in Figure D.2 the function, $\Xi(t)=\omega+\dot{\phi}(t)$, a further interpretation of $\dot{\phi}(t)$ is apparent. The instantaneous frequency of $x(t)$ may, for many practical $x(t)$, appear in a band around center frequency, ω, whose width is a function of the range of $\phi(t)$. The width of this band may be thought of as a measure of the bandwidth of $x(t)$. Clearly, scaling $\phi(t)$ (and therefore $\dot{\phi}(t)$) has the effect of similarly scaling the width of this band. An alternative to this measure of bandwidth is discussed in Section 5.5.

REFERENCES

[1] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*. 2nd ed., New York: Springer-Verlag, 1972.

[2] J. V. Tobias, *Foundations of Modern Auditory Theory*. New York: Academic Press, 1970.

[3] M. R. Schroeder, "Vocoders: Analysis and Synthesis of Speech," *Proc. IEEE*, vol 54, no 5, pp 720-734, May 1966.

[4] H. Helmholtz, *On the Sensation of Tone*. New York: Dover, 1954.

[5] G. von Bekesy, *Experiments in Hearing*. New York: McGraw Hill, 1960.

[6] C. L. Searle, et. al., "Phoneme Recognition Based on Human Audition," Manuscript received from the author, who is with Queen's University, Kingston, Ontario. October 1977.

[7] M. W. Callahan, "Acoustic Signal Processing Based on the Short-Time Spectrum," Ph.D. Thesis, Computer Sci. Dept., Univ. of Utah, Tech. Rept. no. UTEC-CSc-76-209, March 1976.

[8] J. L. Flanagan and R. M. Golden, "Phase Vocoder," *Bell Syst. Tech J.*, vol 45, pp 1493-1509, November 1946.

[9] R. W. Schafer and L. R. Rabiner, "Design and Simulation of a Speech Analysis-Synthesis System Based on Short-Time Fourier Analysis," *IEEE Trans. Audio Electroacoust.*, vol AU-21, pp 165-174, June 1973.

[10] M. R. Portnoff, "Implementation of the Digital Phase Vocoder using the Fast Fourier Transform," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol ASSP-24, no 3, pp 243-248, June 1976.

[11] M. R. Portnoff, "Time-scale Modification of Speech Based on Short-Time Fourier Analysis," Ph.D. dissertation, Dept. of Electrical Engineering and Computer Science, M.I.T., 1978.

[12] G. Gambardella, "Properties of Short-Time Spectral Analysis Performed by the Peripheral Auditory System," Int'l. Congr. Cybernetics 1st., London: Gordon and Breach, 1970.

[13] J. B. Allen, "Short Term Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform," IEEE Trans. Acoust., Speech, Signal Processing, vol ASSP-25, no 3, June 1977.

[14] A. Papoulis, The Fourier Integral and its Applications. New York: McGraw-Hill, 1962.

[15] J. B. Allen and L. R. Rabiner, "A Unified Approach to Short-Time Fourier Analysis and Synthesis," Proc. IEEE, vol 65, no 11, pp 1558-1564, November 1977.

[16] M. J. Lighthill, Fourier Analysis and Generalized Functions. London: Cambridge University Press, 1958.

[17] G. Gambardella, "Time Scaling and Short-Time Spectral Analysis," J. Acoust. Soc. Amer., vol 44, no 6, pp 1745-1747, 1968.

[18] G. Gambardella, "A Contribution to the Theory of Short-Time Spectral Analysis with Nonuniform Bandwidth Filters," IEEE Trans. Circuit Theory, vol CT-18, no 4, pp 455-460, July 1971.

[19] J. T. Kajiya, "Toward a Mathematical Theory of Perception," Ph.D. dissertation, Dept. of Computer Science, Univ. of Utah, 1979.

[20] A. V. Oppenheim, et. al., "Computation of Spectra with Unequal Resolution Using the Fast Fourier Transform," Proc. IEEE, vol 59, no 2, pp 299-301, February 1971.

[21] H. D. Helms, "Power Spectra Obtained from Exponentially Increasing Spacings of Sampling Positions and Frequencies," IEEE Trans. Acoust., Speech and Signal Processing, vol ASSP-24, no 1, February 1976.

[22] E. O. Brigham, The Fast Fourier Transform. Englewood Cliffs, NJ: Prentice Hall, 1974, sec 13.4, pp 217-221.

[23] T. G. Stockham, Jr., "High-speed Convolution and Correlation," AFIPS Proc., vol 28, pp 229-233, 1966 Spring Joint Computer Conf., Washington, D.C.: Spartan, 1966.

[24] L. R.Rabiner and R. E. Crochiere, "Optimum FIR Digital Filter Implementations for Decimation, Interpolation, and Narrow-Band Filtering," IEEE Trans. Acoust., Speech and Signal Processing, vol ASSP-23, no 5, pp 444-456, October 1975.

[25] R. W. Schafer and L. R. Rabiner, "A Digital Signal Processing Approach to Interpolation," Proc IEEE, vol 61, no 6, pp 692-702, June 1973.

[26] J. M. Kates, "Constant-Q Analysis Using the Chirp z-transform," Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing, Washington, D.C., April 2-4, 1979, pp 314-317.

[27] L. R. Rabiner, R. W. Schafer and C. M. Rader, "The Chirp z-transform Algorithm and its Application," Bell Sys. Tech. J., vol 48, pp 1249-1292, May-June 1969.

[28] T. L. Petersen, "Acoustic Signal Processing in the Context of a Perceptual Model," Ph.D. dissertation, Dept. of Computer Science, Univ. of Utah, 1979.

[29] H. Goldstein, "Reading and Listening Comprehension at Various Controlled Rates," Ph.D. Dissertation, Teachers College, Columbia University, 1940.

[30] H. Fletcher, Speech and Hearing. Princeton, NJ: Van Nostrand, 1929, pp 291-294.

[31] E. D. Steinburg, "Effects of Distortion on Speech and Music," Electrical Engineer's Handbook, ed. H. Pender and K. McIlwain, New York: Wiley, 1936, pp 932-938.

[32] G. A. Miller and J. C. Licklider, "The Intelligibility of Interrupted Speech," J. Acoust. Soc. Amer., vol 22, no 3, pp 167-173, March 1950.

[33] W. D. Garvey, "An Experimental Investigation of the Intelligibility of Speeded Speech," Ph.D. Dissertation, Univ. Virginia, 1951.

[34] G. Fairbanks, et. al., "Method for Time of Frequency Compression-Expansion of Speech," Trans. IRE, Professional Group on Audio, vol AU2(1), pp 7-12, 1954.

[35] C. R. Patisaul and J. C. Hammett, "Time-Frequency Experiment in Speech Analysis and Synthesis," J. Acoust. Soc. Amer., vol 58, no 6, pp 1296-1307, December 1975.

[36] R. E. Kahn and J. B. Thomas, "Some Bandwidth Properties of Simultaneous Amplitude and Angle Modulation," IEEE Trans Inf. Theory, vol IT-11, no 4, pp 516-520, October 1965.